Random Variables

2.1. Random Variables

It frequently occurs that in performing an experiment we are mainly interested in some functions of the outcome as opposed to the outcome itself. For instance, in tossing dice we are often interested in the sum of the two dice and are not really concerned about the actual outcome. That is, we may be interested in knowing that the sum is seven and not be concerned over whether the actual outcome was (1, 6) or (2, 5) or (3, 4) or (4, 3) or (5, 2) or (6, 1). These quantities of interest, or more formally, these real-valued functions defined on the sample space, are known as *random variables*.

Since the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to the possible values of the random variable.

Example 2.1 Letting X denote the random variable that is defined as the sum of two fair dice; then

$$P\{X=2\} = P\{(1,1)\} = \tfrac{1}{36},$$

$$P\{X=3\} = P\{(1,2),(2,1)\} = \tfrac{2}{36},$$

$$P\{X=4\} = P\{(1,3),(2,2),(3,1)\} = \tfrac{3}{36},$$

$$P\{X=5\} = P\{(1,4),(2,3),(3,2),(4,1)\} = \tfrac{4}{36},$$

$$P\{X=6\} = P\{(1,5),(2,4),(3,3),(4,2),(5,1)\} = \tfrac{5}{36},$$

$$P\{X=7\} = P\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} = \tfrac{6}{36},$$

$$P\{X=8\} = P\{(2,6),(3,5),(4,4),(5,3),(6,2)\} = \tfrac{5}{36},$$

$$P\{X=9\} = P\{(3,6),(4,5),(5,4),(6,3)\} = \tfrac{4}{36},$$

$$P\{X=10\}=P\{(4,6),(5,5),(6,4)\}=\tfrac{3}{36},$$
$$P\{X=11\}=P\{(5,6),(6,5)\}=\tfrac{2}{36},$$
$$P\{X=12\}=P\{(6,6)\}=\tfrac{1}{36} \qquad (2.1)$$

In other words, the random variable X can take on any integral value between two and twelve, and the probability that it takes on each value is given by Equation (2.1). Since X must take on one of the values two through twelve, we must have that

$$1=P\left\{\bigcup_{i=2}^{12}\{X=n\}\right\}=\sum_{n=2}^{12}P\{X=n\}$$

which may be checked from Equation (2.1). ■

Example 2.2 For a second example, suppose that our experiment consists of tossing two fair coins. Letting Y denote the number of heads appearing, then Y is a random variable taking on one of the values 0, 1, 2 with respective probabilities

$$P\{Y=0\}=P\{(T,T)\}=\tfrac{1}{4},$$
$$P\{Y=1\}=P\{(T,H),(H,T)\}=\tfrac{2}{4},$$
$$P\{Y=2\}=P\{(H,H)\}=\tfrac{1}{4}$$

Of course, $P\{Y=0\}+P\{Y=1\}+P\{Y=2\}=1$. ■

Example 2.3 Suppose that we toss a coin having a probability p of coming up heads, until the first head appears. Letting N denote the number of flips required, then assuming that the outcome of successive flips are independent, N is a random variable taking on one of the values $1, 2, 3, \dots$, with respective probabilities

$$P\{N=1\}=P\{H\}=p,$$
$$P\{N=2\}=P\{(T,H)\}=(1-p)p,$$
$$P\{N=3\}=P\{(T,T,H)\}=(1-p)^2p,$$
$$\vdots$$
$$P\{N=n\}=P\{(\underbrace{T,T,\dots,T}_{n-1},H)\}=(1-p)^{n-1}p, \qquad n\geqslant 1$$

As a check, note that

$$
\begin{aligned}
P\left(\bigcup_{n=1}^{\infty}\{N=n\}\right) &= \sum_{n=1}^{\infty} P\{N=n\} \\
&= p\sum_{n=1}^{\infty}(1-p)^{n-1} \\
&= \frac{p}{1-(1-p)} \\
&= 1 \quad \blacksquare
\end{aligned}
$$

Example 2.4 Suppose that our experiment consists of seeing how long a battery can operate before wearing down. Suppose also that we are not primarily interested in the actual lifetime of the battery but are concerned only about whether or not the battery lasts at least two years. In this case, we may define the random variable I by

$$
I = \begin{cases} 1, & \text{if the lifetime of battery is two or more years} \\ 0, & \text{otherwise} \end{cases}
$$

If E denotes the event that the battery lasts two or more years, then the random variable I is known as the *indicator* random variable for event E. (Note that I equals 1 or 0 depending on whether or not E occurs.) $\blacksquare$

Example 2.5 Suppose that independent trials, each of which results in any of m possible outcomes with respective probabilities $p_1, \dots, p_m$, $\sum_{i=1}^m p_i = 1$, are continually performed. Let X denote the number of trials needed until each outcome has occurred at least once.

Rather than directly considering $P\{X = n\}$ we will first determine $P\{X > n\}$, the probability that at least one of the outcomes has not yet occurred after n trials. Letting A_i denote the event that outcome i has not yet occurred after the first n trials, $i = 1, \dots, m$, then

$$
\begin{aligned}
P\{X > n\} &= P\left(\bigcup_{i=1}^{m} A_i\right) \\
&= \sum_{i=1}^{m} P(A_i) - \sum\sum_{i<j} P(A_i A_j) \\
&\quad + \sum\sum\sum_{i<j<k} P(A_i A_j A_k) - \cdots + (-1)^{m+1} P(A_1 \cdots A_m)
\end{aligned}
$$

Now, $P(A_i)$ is the probability that each of the first n trials results in a non-i outcome, and so by independence

$$P(A_i) = (1 - p_i)^n$$

Similarly, $P(A_i A_j)$ is the probability that the first n trials all result in a non-i and non-j outcome, and so

$$P(A_i A_j) = (1 - p_i - p_j)^n$$

As all of the other probabilities are similar, we see that

$$\begin{aligned} P\{X > n\} = \sum_{i=1}^{m} (1 - p_i)^n - \sum\sum_{i<j} (1 - p_i - p_j)^n \\ + \sum\sum\sum_{i<j<k} (1 - p_i - p_j - p_k)^n - \cdots \end{aligned}$$

Since $P\{X = n\} = P\{X > n - 1\} - P\{X > n\}$, we see, upon using the algebraic identity $(1 - a)^{n-1} - (1 - a)^n = a(1 - a)^{n-1}$, that

$$\begin{aligned} P\{X = n\} = \sum_{i=1}^{m} p_i (1 - p_i)^{n-1} - \sum\sum_{i<j} (p_i + p_j)(1 - p_i - p_j)^{n-1} \\ + \sum\sum\sum_{i<j<k} (p_i + p_j + p_k)(1 - p_i - p_j - p_k)^{n-1} - \cdots \quad \blacksquare \end{aligned}$$

In all of the preceding examples, the random variables of interest took on either a finite or a countable number of possible values.* Such random variables are called *discrete*. However, there also exist random variables that take on a continuum of possible values. These are known as *continuous* random variables. One example is the random variable denoting the lifetime of a car, when the car's lifetime is assumed to take on any value in some interval (a, b).

The *cumulative distribution function* (cdf) (or more simply the *distribution function*) $F(\cdot)$ of the random variable X is defined for any real number b, $-\infty < b < \infty$, by

$$F(b) = P\{X \leqslant b\}$$

In words, $F(b)$ denotes the probability that the random variable X takes on a value that is less than or equal to b. Some properties of the cdf F are

(i) $F(b)$ is a nondecreasing function of b,

*A set is countable if its elements can be put in a one-to-one correspondence with the sequence of positive integers.

(ii) $\lim_{b\to\infty} F(b) = F(\infty) = 1$,
(iii) $\lim_{b\to-\infty} F(b) = F(-\infty) = 0$.

Property (i) follows since for $a < b$ the event $\{X \leqslant a\}$ is contained in the event $\{X \leqslant b\}$, and so it must have a smaller probability. Properties (ii) and (iii) follow since X must take on some finite value.

All probability questions about X can be answered in terms of the cdf $F(\cdot)$. For example,

$$P\{a < X \leqslant b\} = F(b) - F(a) \qquad \text{for all } a < b$$

This follows since we may calculate $P\{a < X \leqslant b\}$ by first computing the probability that $X \leqslant b$ [that is, $F(b)$] and then subtracting from this the probability that $X \leqslant a$ [that is, $F(a)$].

If we desire the probability that X is strictly smaller than b, we may calculate this probability by

$$\begin{aligned} P\{X < b\} &= \lim_{h\to 0^+} P\{X \leqslant b - h\} \\ &= \lim_{h\to 0^+} F(b-h) \end{aligned}$$

where $\lim_{h\to 0^+}$ means that we are taking the limit as h decreases to 0. Note that $P\{X < b\}$ does not necessarily equal $F(b)$ since $F(b)$ also includes the probability that X equals b.

2.2. Discrete Random Variables

As was previously mentioned, a random variable that can take on at most a countable number of possible values is said to be *discrete*. For a discrete random variable X, we define the *probability mass function* $p(a)$ of X by

$$p(a) = P\{X = a\}$$

The probability mass function $p(a)$ is positive for at most a countable number of values of a. That is, if X must assume one of the values $x_1, x_2, \ldots$, then

$$\begin{aligned} p(x_i) &> 0, \qquad i = 1, 2, \ldots \\ p(x) &= 0, \qquad \text{all other values of } x \end{aligned}$$

Since X must take on one of the values x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

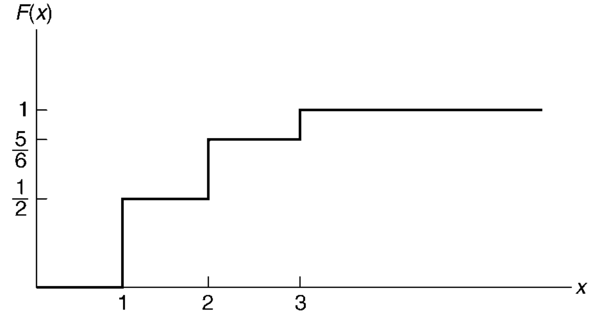

Figure 2.1. Graph of $F(x)$.

The cumulative distribution function F can be expressed in terms of $p(a)$ by

$$F(a) = \sum_{\text{all } x_i \leqslant a} p(x_i)$$

For instance, suppose X has a probability mass function given by

$$p(1) = \tfrac{1}{2}, \qquad p(2) = \tfrac{1}{3}, \qquad p(3) = \tfrac{1}{6}$$

then, the cumulative distribution function F of X is given by

$$F(a) = \begin{cases} 0, & a < 1 \\ \frac{1}{2}, & 1 \leqslant a < 2 \\ \frac{5}{6}, & 2 \leqslant a < 3 \\ 1, & 3 \leqslant a \end{cases}$$

This is graphically presented in Figure 2.1.

Discrete random variables are often classified according to their probability mass functions. We now consider some of these random variables.

2.2.1. The Bernoulli Random Variable

Suppose that a trial, or an experiment, whose outcome can be classified as either a "success" or as a "failure" is performed. If we let X equal 1 if the outcome is a success and 0 if it is a failure, then the probability mass function of X is given by

$$\begin{aligned} p(0) &= P\{X = 0\} = 1 - p, \\ p(1) &= P\{X = 1\} = p \end{aligned} \tag{2.2}$$

where p, $0 \leqslant p \leqslant 1$, is the probability that the trial is a "success."

A random variable X is said to be a *Bernoulli* random variable if its probability mass function is given by Equation (2.2) for some $p \in (0, 1)$.

2.2.2. The Binomial Random Variable

Suppose that n independent trials, each of which results in a "success" with probability p and in a "failure" with probability $1 - p$, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a *binomial* random variable with parameters (n, p).

The probability mass function of a binomial random variable having parameters (n, p) is given by

$$p(i) = \binom{n}{i} p^i (1 - p)^{n-i}, \qquad i = 0, 1, \ldots, n \tag{2.3}$$

where

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

equals the number of different groups of i objects that can be chosen from a set of n objects. The validity of Equation (2.3) may be verified by first noting that the probability of any particular sequence of the n outcomes containing i successes and $n - i$ failures is, by the assumed independence of trials, $p^i (1 - p)^{n-i}$. Equation (2.3) then follows since there are $\binom{n}{i}$ different sequences of the n outcomes leading to i successes and $n - i$ failures. For instance, if $n = 3, i = 2$, then there are $\binom{3}{2} = 3$ ways in which the three trials can result in two successes. Namely, any one of the three outcomes (s, s, f), (s, f, s), (f, s, s), where the outcome (s, s, f) means that the first two trials are successes and the third a failure. Since each of the three outcomes (s, s, f), (s, f, s), (f, s, s) has a probability $p^2(1 - p)$ of occurring the desired probability is thus $\binom{3}{2} p^2 (1 - p)$.

Note that, by the binomial theorem, the probabilities sum to one, that is,

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1 - p)^{n-i} = \big(p + (1 - p)\big)^n = 1$$

Example 2.6 Four fair coins are flipped. If the outcomes are assumed independent, what is the probability that two heads and two tails are obtained?

Solution: Letting X equal the number of heads ("successes") that appear, then X is a binomial random variable with parameters $(n = 4, p = \frac{1}{2})$.

Hence, by Equation (2.3),

$$P\{X = 2\} = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{3}{8} \quad \blacksquare$$

Example 2.7 It is known that any item produced by a certain machine will be defective with probability 0.1, independently of any other item. What is the probability that in a sample of three items, at most one will be defective?

Solution: If X is the number of defective items in the sample, then X is a binomial random variable with parameters $(3, 0.1)$. Hence, the desired probability is given by

$$P\{X = 0\} + P\{X = 1\} = \binom{3}{0}(0.1)^0(0.9)^3 + \binom{3}{1}(0.1)^1(0.9)^2 = 0.972 \quad \blacksquare$$

Example 2.8 Suppose that an airplane engine will fail, when in flight, with probability $1 - p$ independently from engine to engine; suppose that the airplane will make a successful flight if at least 50 percent of its engines remain operative. For what values of p is a four-engine plane preferable to a two-engine plane?

Solution: Because each engine is assumed to fail or function independently of what happens with the other engines, it follows that the number of engines remaining operative is a binomial random variable. Hence, the probability that a four-engine plane makes a successful flight is

$$\begin{aligned}
&\binom{4}{2}p^2(1-p)^2 + \binom{4}{3}p^3(1-p) + \binom{4}{4}p^4(1-p)^0 \\
&\quad = 6p^2(1-p)^2 + 4p^3(1-p) + p^4
\end{aligned}$$

whereas the corresponding probability for a two-engine plane is

$$\binom{2}{1}p(1-p) + \binom{2}{2}p^2 = 2p(1-p) + p^2$$

Hence the four-engine plane is safer if

$$6p^2(1-p)^2 + 4p^3(1-p) + p^4 \geqslant 2p(1-p) + p^2$$

or equivalently if

$$6p(1-p)^2 + 4p^2(1-p) + p^3 \geqslant 2 - p$$

which simplifies to

$$3p^3 - 8p^2 + 7p - 2 \geqslant 0 \quad \text{or} \quad (p-1)^2(3p-2) \geqslant 0$$

which is equivalent to

$$3p - 2 \geqslant 0 \quad \text{or} \quad p \geqslant \tfrac{2}{3}$$

Hence, the four-engine plane is safer when the engine success probability is at least as large as $\frac{2}{3}$, whereas the two-engine plane is safer when this probability falls below $\frac{2}{3}$. ■

Example 2.9 Suppose that a particular trait of a person (such as eye color or left handedness) is classified on the basis of one pair of genes and suppose that d represents a dominant gene and r a recessive gene. Thus a person with dd genes is pure dominance, one with rr is pure recessive, and one with rd is hybrid. The pure dominance and the hybrid are alike in appearance. Children receive one gene from each parent. If, with respect to a particular trait, two hybrid parents have a total of four children, what is the probability that exactly three of the four children have the outward appearance of the dominant gene?

Solution: If we assume that each child is equally likely to inherit either of two genes from each parent, the probabilities that the child of two hybrid parents will have dd, rr, or rd pairs of genes are, respectively, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{2}$. Hence, because an offspring will have the outward appearance of the dominant gene if its gene pair is either dd or rd, it follows that the number of such children is binomially distributed with parameters $(4, \frac{3}{4})$. Thus the desired probability is

$$\binom{4}{3}\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)^1 = \frac{27}{64} \quad ■$$

Remark on Terminology If X is a binomial random variable with parameters (n, p), then we say that X has a binomial distribution with parameters (n, p).

2.2.3. The Geometric Random Variable

Suppose that independent trials, each having probability p of being a success, are performed until a success occurs. If we let X be the number of trials required until the first success, then X is said to be a *geometric* random variable with parameter p. Its probability mass function is given by

$$p(n) = P\{X = n\} = (1 - p)^{n-1} p, \qquad n = 1, 2, \ldots \tag{2.4}$$

Equation (2.4) follows since in order for X to equal n it is necessary and sufficient that the first $n - 1$ trials be failures and the nth trial a success. Equation (2.4) follows since the outcomes of the successive trials are assumed to be independent.

To check that $p(n)$ is a probability mass function, we note that

$$\sum_{n=1}^{\infty} p(n) = p \sum_{n=1}^{\infty} (1 - p)^{n-1} = 1$$

2.2.4. The Poisson Random Variable

A random variable X, taking on one of the values $0, 1, 2, \ldots$, is said to be a *Poisson* random variable with parameter λ, if for some $\lambda > 0$,

$$p(i) = P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, \ldots \tag{2.5}$$

Equation (2.5) defines a probability mass function since

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1$$

The Poisson random variable has a wide range of applications in a diverse number of areas, as will be seen in Chapter 5.

An important property of the Poisson random variable is that it may be used to approximate a binomial random variable when the binomial parameter n is large and p is small. To see this, suppose that X is a binomial random variable with parameters (n, p), and let $\lambda = np$. Then

$$\begin{aligned}
P\{X = i\} &= \frac{n!}{(n - i)!\, i!} p^i (1 - p)^{n-i} \\
&= \frac{n!}{(n - i)!\, i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\
&= \frac{n(n - 1) \cdots (n - i + 1)}{n^i} \frac{\lambda^i}{i!} \frac{(1 - \lambda/n)^n}{(1 - \lambda/n)^i}
\end{aligned}$$

Now, for n large and p small

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \qquad \frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1, \qquad \left(1-\frac{\lambda}{n}\right)^i \approx 1$$

Hence, for n large and p small,

$$P\{X=i\} \approx e^{-\lambda}\frac{\lambda^i}{i!}$$

Example 2.10 Suppose that the number of typographical errors on a single page of this book has a Poisson distribution with parameter $\lambda = 1$. Calculate the probability that there is at least one error on this page.

Solution:

$$P\{X \geqslant 1\} = 1 - P\{X=0\} = 1 - e^{-1} \approx 0.633 \quad \blacksquare$$

Example 2.11 If the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\lambda = 3$, what is the probability that no accidents occur today?

Solution:

$$P\{X=0\} = e^{-3} \approx 0.05 \quad \blacksquare$$

Example 2.12 Consider an experiment that consists of counting the number of α-particles given off in a one-second interval by one gram of radioactive material. If we know from past experience that, on the average, 3.2 such α-particles are given off, what is a good approximation to the probability that no more than two α-particles will appear?

Solution: If we think of the gram of radioactive material as consisting of a large number n of atoms each of which has probability $3.2/n$ of disintegrating and sending off an α-particle during the second considered, then we see that, to a very close approximation, the number of α-particles given off will be a Poisson random variable with parameter $\lambda = 3.2$. Hence the desired probability is

$$P\{X \leqslant 2\} = e^{-3.2} + 3.2e^{-3.2} + \frac{(3.2)^2}{2}e^{-3.2} \approx 0.382 \quad \blacksquare$$

2.3. Continuous Random Variables

In this section, we shall concern ourselves with random variables whose set of possible values is uncountable. Let X be such a random variable. We say that X is a *continuous* random variable if there exists a nonnegative function $f(x)$, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers

$$P\{X \in B\} = \int_B f(x)\,dx \tag{2.6}$$

The function $f(x)$ is called the *probability density function* of the random variable X.

In words, Equation (2.6) states that the probability that X will be in B may be obtained by integrating the probability density function over the set B. Since X must assume some value, $f(x)$ must satisfy

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx$$

All probability statements about X can be answered in terms of $f(x)$. For instance, letting $B = [a, b]$, we obtain from Equation (2.6) that

$$P\{a \leqslant X \leqslant b\} = \int_a^b f(x)\,dx \tag{2.7}$$

If we let $a = b$ in the preceding, then

$$P\{X = a\} = \int_a^a f(x)\,dx = 0$$

In words, this equation states that the probability that a continuous random variable will assume any *particular* value is zero.

The relationship between the cumulative distribution $F(\cdot)$ and the probability density $f(\cdot)$ is expressed by

$$F(a) = P\{X \in (-\infty, a]\} = \int_{-\infty}^{a} f(x)\,dx$$

Differentiating both sides of the preceding yields

$$\frac{d}{da}F(a) = f(a)$$

That is, the density is the derivative of the cumulative distribution function. A somewhat more intuitive interpretation of the density function may be obtained from Equation (2.7) as follows:

$$P\left\{a - \frac{\varepsilon}{2} \leqslant X \leqslant a + \frac{\varepsilon}{2}\right\} = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)\,dx \approx \varepsilon f(a)$$

when ε is small. In other words, the probability that X will be contained in an interval of length ε around the point a is approximately $\varepsilon f(a)$. From this, we see that $f(a)$ is a measure of how likely it is that the random variable will be near a.

There are several important continuous random variables that appear frequently in probability theory. The remainder of this section is devoted to a study of certain of these random variables.

2.3.1. The Uniform Random Variable

A random variable is said to be uniformly distributed over the interval (0, 1) if its probability density function is given by

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Note that the preceding is a density function since $f(x) \geqslant 0$ and

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 dx = 1$$

Since $f(x) > 0$ only when $x \in (0, 1)$, it follows that X must assume a value in $(0, 1)$. Also, since $f(x)$ is constant for $x \in (0, 1)$, X is just as likely to be "near" any value in (0, 1) as any other value. To check this, note that, for any $0 < a < b < 1$,

$$P\{a \leqslant X \leqslant b\} = \int_a^b f(x)\,dx = b - a$$

In other words, the probability that X is in any particular subinterval of (0, 1) equals the length of that subinterval.

In general, we say that X is a uniform random variable on the interval (α, β) if its probability density function is given by

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha}, & \text{if } \alpha < x < \beta \\ 0, & \text{otherwise} \end{cases} \tag{2.8}$$

Example 2.13 Calculate the cumulative distribution function of a random variable uniformly distributed over (α, β).

Solution: Since $F(a) = \int_{-\infty}^{a} f(x)\,dx$, we obtain from Equation (2.8) that

$$F(a) = \begin{cases} 0, & a \leqslant \alpha \\ \dfrac{a-\alpha}{\beta-\alpha}, & \alpha < a < \beta \\ 1, & a \geqslant \beta \end{cases} \quad \blacksquare$$

Example 2.14 If X is uniformly distributed over (0, 10), calculate the probability that (a) $X < 3$, (b) $X > 7$, (c) $1 < X < 6$.

Solution:

$$P\{X < 3\} = \frac{\int_0^3 dx}{10} = \frac{3}{10},$$

$$P\{X > 7\} = \frac{\int_7^{10} dx}{10} = \frac{3}{10},$$

$$P\{1 < X < 6\} = \frac{\int_1^6 dx}{10} = \frac{1}{2} \quad \blacksquare$$

2.3.2. Exponential Random Variables

A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0 \end{cases}$$

is said to be an exponential random variable with parameter λ. These random variables will be extensively studied in Chapter 5, so we will content ourselves here with just calculating the cumulative distribution function F:

$$F(a) = \int_0^a \lambda e^{-\lambda x} = 1 - e^{-\lambda a}, \qquad a \geqslant 0$$

Note that $F(\infty) = \int_0^\infty \lambda e^{-\lambda x}\,dx = 1$, as, of course, it must.

2.3.3. Gamma Random Variables

A continuous random variable whose density is given by

$$f(x) = \begin{cases} \dfrac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0 \end{cases}$$

for some $\lambda > 0$, $\alpha > 0$ is said to be a gamma random variable with parameters α, λ. The quantity $\Gamma(\alpha)$ is called the gamma function and is defined by

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

It is easy to show by induction that for integral α, say, $\alpha = n$,

$$\Gamma(n) = (n-1)!$$

2.3.4. Normal Random Variables

We say that X is a normal random variable (or simply that X is normally distributed) with parameters μ and σ^2 if the density of X is given by

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

This density function is a bell-shaped curve that is symmetric around μ (see Figure 2.2).

An important fact about normal random variables is that if X is normally distributed with parameters μ and σ^2 then $Y = \alpha X + \beta$ is normally distributed with parameters $\alpha\mu + \beta$ and $\alpha^2\sigma^2$. To prove this, suppose first that $\alpha > 0$ and note that

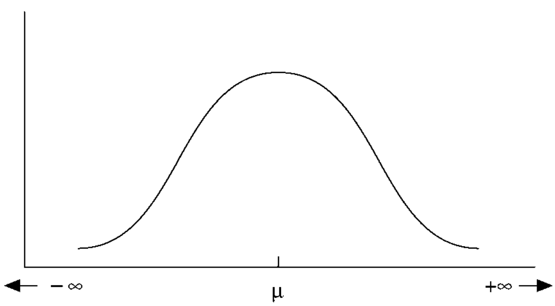

Figure 2.2. Normal density function.

$F_Y(\cdot)$* the cumulative distribution function of the random variable Y is given by

$$\begin{aligned}
F_Y(a) &= P\{Y \leqslant a\} \\
&= P\{\alpha X + \beta \leqslant a\} \\
&= P\left\{X \leqslant \frac{a-\beta}{\alpha}\right\} \\
&= F_X\left(\frac{a-\beta}{\alpha}\right) \\
&= \int_{-\infty}^{(a-\beta)/\alpha} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx \\
&= \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\,\alpha\sigma} \exp\left\{\frac{-(v-(\alpha\mu+\beta))^2}{2\alpha^2\sigma^2}\right\} dv
\end{aligned} \tag{2.9}$$

where the last equality is obtained by the change in variables $v = \alpha x + \beta$. However, since $F_Y(a) = \int_{-\infty}^{a} f_Y(v)\, dv$, it follows from Equation (2.9) that the probability density function $f_Y(\cdot)$ is given by

$$f_Y(v) = \frac{1}{\sqrt{2\pi}\alpha\sigma} \exp\left\{\frac{-(v-(\alpha\mu+\beta))^2}{2(\alpha\sigma)^2}\right\}, \qquad -\infty < v < \infty$$

Hence, Y is normally distributed with parameters $\alpha\mu + \beta$ and $(\alpha\sigma)^2$. A similar result is also true when $\alpha < 0$.

One implication of the preceding result is that if X is normally distributed with parameters μ and σ^2 then $Y = (X - \mu)/\sigma$ is normally distributed with parameters 0 and 1. Such a random variable Y is said to have the *standard* or *unit* normal distribution.

2.4. Expectation of a Random Variable

2.4.1. The Discrete Case

If X is a discrete random variable having a probability mass function $p(x)$, then the *expected value* of X is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

*When there is more than one random variable under consideration, we shall denote the cumulative distribution function of a random variable Z by $F_z(\cdot)$. Similarly, we shall denote the density of Z by $f_z(\cdot)$.

In other words, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes that value. For example, if the probability mass function of X is given by

$$p(1) = \tfrac{1}{2} = p(2)$$

then

$$E[X] = 1(\tfrac{1}{2}) + 2(\tfrac{1}{2}) = \tfrac{3}{2}$$

is just an ordinary average of the two possible values 1 and 2 that X can assume. On the other hand, if

$$p(1) = \tfrac{1}{3}, \qquad p(2) = \tfrac{2}{3}$$

then

$$E[X] = 1(\tfrac{1}{3}) + 2(\tfrac{2}{3}) = \tfrac{5}{3}$$

is a weighted average of the two possible values 1 and 2 where the value 2 is given twice as much weight as the value 1 since $p(2) = 2p(1)$.

Example 2.15 Find $E[X]$ where X is the outcome when we roll a fair die.

Solution: Since $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = \frac{1}{6}$, we obtain

$$E[X] = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + 3(\tfrac{1}{6}) + 4(\tfrac{1}{6}) + 5(\tfrac{1}{6}) + 6(\tfrac{1}{6}) = \tfrac{7}{2} \quad \blacksquare$$

Example 2.16 (Expectation of a Bernoulli Random Variable) Calculate $E[X]$ when X is a Bernoulli random variable with parameter p.

Solution: Since $p(0) = 1 - p,\ p(1) = p$, we have

$$E[X] = 0(1 - p) + 1(p) = p$$

Thus, the expected number of successes in a single trial is just the probability that the trial will be a success. $\blacksquare$

Example 2.17 (Expectation of a Binomial Random Variable) Calculate $E[X]$ when X is binomially distributed with parameters n and p.

Solution:

$$\begin{aligned}
E[X] &= \sum_{i=0}^{n} i p(i) \\
&= \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} \\
&= \sum_{i=1}^{n} \frac{in!}{(n-i)!\,i!} p^i (1-p)^{n-i} \\
&= \sum_{i=1}^{n} \frac{n!}{(n-i)!(i-1)!} p^i (1-p)^{n-i} \\
&= np \sum_{i=1}^{n} \frac{(n-1)!}{(n-i)!(i-1)!} p^{i-1} (1-p)^{n-i} \\
&= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} \\
&= np[p + (1-p)]^{n-1} \\
&= np
\end{aligned}$$

where the second from the last equality follows by letting $k = i - 1$. Thus, the expected number of successes in n independent trials is n multiplied by the probability that a trial results in a success. ■

Example 2.18 (Expectation of a Geometric Random Variable) Calculate the expectation of a geometric random variable having parameter p.

Solution: By Equation (2.4), we have

$$\begin{aligned}
E[X] &= \sum_{n=1}^{\infty} np(1-p)^{n-1} \\
&= p \sum_{n=1}^{\infty} nq^{n-1}
\end{aligned}$$

where $q = 1 - p$,

$$\begin{aligned}
E[X] &= p \sum_{n=1}^{\infty} \frac{d}{dq}(q^n) \\
&= p \frac{d}{dq} \left(\sum_{n=1}^{\infty} q^n \right)
\end{aligned}$$

$$
\begin{aligned}
&= p\frac{d}{dq}\left(\frac{q}{1-q}\right) \\
&= \frac{p}{(1-q)^2} \\
&= \frac{1}{p}
\end{aligned}
$$

In words, the expected number of independent trials we need to perform until we attain our first success equals the reciprocal of the probability that any one trial results in a success. ■

Example 2.19 (Expectation of a Poisson Random Variable) Calculate $E[X]$ if X is a Poisson random variable with parameter λ.

Solution: From Equation (2.5), we have

$$
\begin{aligned}
E[X] &= \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} \\
&= \sum_{i=1}^{\infty} \frac{e^{-\lambda}\lambda^i}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}
$$

where we have used the identity $\sum_{k=0}^{\infty} \lambda^k/k! = e^{\lambda}$. ■

2.4.2. The Continuous Case

We may also define the expected value of a continuous random variable. This is done as follows. If X is a continuous random variable having a probability density function $f(x)$, then the expected value of X is defined by

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

Example 2.20 (Expectation of a Uniform Random Variable) Calculate the expectation of a random variable uniformly distributed over (α, β).

Solution: From Equation (2.8) we have

$$\begin{aligned} E[X] &= \int_\alpha^\beta \frac{x}{\beta - \alpha}\, dx \\ &= \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} \\ &= \frac{\beta + \alpha}{2} \end{aligned}$$

In other words, the expected value of a random variable uniformly distributed over the interval (α, β) is just the midpoint of the interval. ■

Example 2.21 (Expectation of an Exponential Random Variable) Let X be exponentially distributed with parameter λ. Calculate $E[X]$.

Solution:

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x}\, dx$$

Integrating by parts yields

$$\begin{aligned} E[X] &= -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\, dx \\ &= 0 - \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty \\ &= \frac{1}{\lambda} \quad ■ \end{aligned}$$

Example 2.22 (Expectation of a Normal Random Variable) Calculate $E[X]$ when X is normally distributed with parameters μ and σ^2.

Solution:

$$E[X] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^\infty x e^{-(x-\mu)^2/2\sigma^2}\, dx$$

Writing x as $(x - \mu) + \mu$ yields

$$E[X] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^\infty (x - \mu) e^{-(x-\mu)^2/2\sigma^2}\, dx + \mu \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^\infty e^{-(x-\mu)^2/2\sigma^2}\, dx$$

Letting $y = x - \mu$ leads to

$$E[X] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} y e^{-y^2/2\sigma^2} \, dy + \mu \int_{-\infty}^{\infty} f(x) \, dx$$

where $f(x)$ is the normal density. By symmetry, the first integral must be 0, and so

$$E[X] = \mu \int_{-\infty}^{\infty} f(x) \, dx = \mu \quad \blacksquare$$

2.4.3. Expectation of a Function of a Random Variable

Suppose now that we are given a random variable X and its probability distribution (that is, its probability mass function in the discrete case or its probability density function in the continuous case). Suppose also that we are interested in calculating, not the expected value of X, but the expected value of some function of X, say, $g(X)$. How do we go about doing this? One way is as follows. Since $g(X)$ is itself a random variable, it must have a probability distribution, which should be computable from a knowledge of the distribution of X. Once we have obtained the distribution of $g(X)$, we can then compute $E[g(X)]$ by the definition of the expectation.

Example 2.23 Suppose X has the following probability mass function:

$$p(0) = 0.2, \qquad p(1) = 0.5, \qquad p(2) = 0.3$$

Calculate $E[X^2]$.

Solution: Letting $Y = X^2$, we have that Y is a random variable that can take on one of the values $0^2, 1^2, 2^2$ with respective probabilities

$$\begin{aligned}
p_Y(0) &= P\{Y = 0^2\} = 0.2, \\
p_Y(1) &= P\{Y = 1^2\} = 0.5, \\
p_Y(4) &= P\{Y = 2^2\} = 0.3
\end{aligned}$$

Hence,

$$E[X^2] = E[Y] = 0(0.2) + 1(0.5) + 4(0.3) = 1.7$$

Note that

$$1.7 = E[X^2] \neq (E[X])^2 = 1.21 \quad \blacksquare$$

Example 2.24 Let X be uniformly distributed over (0, 1). Calculate $E[X^3]$.

Solution: Letting $Y = X^3$, we calculate the distribution of Y as follows. For $0 \leqslant a \leqslant 1$,

$$\begin{aligned} F_Y(a) &= P\{Y \leqslant a\} \\ &= P\{X^3 \leqslant a\} \\ &= P\{X \leqslant a^{1/3}\} \\ &= a^{1/3} \end{aligned}$$

where the last equality follows since X is uniformly distributed over (0, 1). By differentiating $F_Y(a)$, we obtain the density of Y, namely,

$$f_Y(a) = \tfrac{1}{3}a^{-2/3}, \qquad 0 \leqslant a \leqslant 1$$

Hence,

$$\begin{aligned} E[X^3] = E[Y] &= \int_{-\infty}^{\infty} a f_Y(a)\, da \\ &= \int_0^1 a\tfrac{1}{3}a^{-2/3}\, da \\ &= \tfrac{1}{3}\int_0^1 a^{1/3}\, da \\ &= \tfrac{1}{3}\tfrac{3}{4}a^{4/3}\Big|_0^1 \\ &= \tfrac{1}{4} \quad \blacksquare \end{aligned}$$

While the foregoing procedure will, in theory, always enable us to compute the expectation of any function of X from a knowledge of the distribution of X, there is, fortunately, an easier way to do this. The following proposition shows how we can calculate the expectation of $g(X)$ without first determining its distribution.

Proposition 2.1 (a) If X is a discrete random variable with probability mass function $p(x)$, then for any real-valued function g,

$$E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$$

(b) If X is a continuous random variable with probability density function $f(x)$, then for any real-valued function g,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \quad \blacksquare$$

Example 2.25 Applying the proposition to Example 2.23 yields

$$E[X^2] = 0^2(0.2) + (1^2)(0.5) + (2^2)(0.3) = 1.7$$

which, of course, checks with the result derived in Example 2.23. $\blacksquare$

Example 2.26 Applying the proposition to Example 2.24 yields

$$\begin{aligned} E[X^3] &= \int_0^1 x^3\, dx \qquad (\text{since } f(x) = 1,\ 0 < x < 1) \\ &= \tfrac{1}{4} \quad \blacksquare \end{aligned}$$

A simple corollary of Proposition 2.1 is the following.

Corollary 2.2 If a and b are constants, then

$$E[aX + b] = aE[X] + b$$

Proof In the discrete case,

$$\begin{aligned} E[aX + b] &= \sum_{x:p(x)>0} (ax + b)p(x) \\ &= a \sum_{x:p(x)>0} xp(x) + b \sum_{x:p(x)>0} p(x) \\ &= aE[X] + b \end{aligned}$$

In the continuous case,

$$\begin{aligned} E[aX + b] &= \int_{-\infty}^{\infty} (ax + b) f(x)\, dx \\ &= a \int_{-\infty}^{\infty} x f(x)\, dx + b \int_{-\infty}^{\infty} f(x)\, dx \\ &= aE[X] + b \quad \blacksquare \end{aligned}$$

The expected value of a random variable X, $E[X]$, is also referred to as the *mean* or the first *moment* of X. The quantity $E[X^n]$, $n \geqslant 1$, is called the nth moment of X. By Proposition 2.1, we note that

$$E[X^n] = \begin{cases} \displaystyle\sum_{x:p(x)>0} x^n p(x), & \text{if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} x^n f(x)\, dx, & \text{if } X \text{ is continuous} \end{cases}$$

Another quantity of interest is the variance of a random variable X, denoted by $\operatorname{Var}(X)$, which is defined by

$$\operatorname{Var}(X) = E\big[(X - E[X])^2\big]$$

Thus, the variance of X measures the expected square of the deviation of X from its expected value.

Example 2.27 (Variance of the Normal Random Variable) Let X be normally distributed with parameters μ and σ^2. Find $\operatorname{Var}(X)$.

Solution: Recalling (see Example 2.22) that $E[X] = \mu$, we have that

$$\begin{aligned} \operatorname{Var}(X) &= E[(X-\mu)^2] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2}\, dx \end{aligned}$$

Substituting $y = (x-\mu)/\sigma$ yields

$$\operatorname{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2}\, dy$$

Integrating by parts $(u = y,\ dv = ye^{-y^2/2}dy)$ gives

$$\begin{aligned} \operatorname{Var}(X) &= \frac{\sigma^2}{\sqrt{2\pi}} \left(-ye^{-y^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-y^2/2}\, dy \right) \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy \\ &= \sigma^2 \end{aligned}$$

Another derivation of $\operatorname{Var}(X)$ will be given in Example 2.42. ■

Suppose that X is continuous with density f, and let $E[X] = \mu$. Then,

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x)\, dx \\
&= \int_{-\infty}^{\infty} x^2 f(x)\, dx - 2\mu \int_{-\infty}^{\infty} x f(x)\, dx + \mu^2 \int_{-\infty}^{\infty} f(x)\, dx \\
&= E[X^2] - 2\mu\mu + \mu^2 \\
&= E[X^2] - \mu^2
\end{aligned}$$

A similar proof holds in the discrete case, and so we obtain the useful identity

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

Example 2.28 Calculate $\mathrm{Var}(X)$ when X represents the outcome when a fair die is rolled.

Solution: As previously noted in Example 2.15, $E[X] = \frac{7}{2}$. Also,

$$E[X^2] = 1\left(\tfrac{1}{6}\right) + 2^2\left(\tfrac{1}{6}\right) + 3^2\left(\tfrac{1}{6}\right) + 4^2\left(\tfrac{1}{6}\right) + 5^2\left(\tfrac{1}{6}\right) + 6^2\left(\tfrac{1}{6}\right) = \left(\tfrac{1}{6}\right)(91)$$

Hence,

$$\mathrm{Var}(X) = \tfrac{91}{6} - \left(\tfrac{7}{2}\right)^2 = \tfrac{35}{12} \quad \blacksquare$$

2.5. Jointly Distributed Random Variables

2.5.1. Joint Distribution Functions

Thus far, we have concerned ourselves with the probability distribution of a single random variable. However, we are often interested in probability statements concerning two or more random variables. To deal with such probabilities, we define, for any two random variables X and Y, the *joint cumulative probability distribution function* of X and Y by

$$F(a, b) = P\{X \leqslant a, Y \leqslant b\}, \qquad -\infty < a, b < \infty$$

The distribution of X can be obtained from the joint distribution of X and Y as follows:

$$\begin{aligned}F_X(a) &= P\{X \leqslant a\}\\ &= P\{X \leqslant a,\ Y < \infty\}\\ &= F(a, \infty)\end{aligned}$$

Similarly, the cumulative distribution function of Y is given by

$$F_Y(b) = P\{Y \leqslant b\} = F(\infty,\ b)$$

In the case where X and Y are both discrete random variables, it is convenient to define the *joint probability mass function* of X and Y by

$$p(x, y) = P\{X = x,\ Y = y\}$$

The probability mass function of X may be obtained from $p(x, y)$ by

$$p_X(x) = \sum_{y:p(x,y)>0} p(x, y)$$

Similarly,

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x, y)$$

We say that X and Y are *jointly continuous* if there exists a function $f(x, y)$, defined for all real x and y, having the property that for all sets A and B of real numbers

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x, y)\,dx\,dy$$

The function $f(x, y)$ is called the *joint probability density function* of X and Y. The probability density of X can be obtained from a knowledge of $f(x, y)$ by the following reasoning:

$$\begin{aligned}P\{X \in A\} &= P\{X \in A,\ Y \in (-\infty, \infty)\}\\ &= \int_{-\infty}^{\infty} \int_A f(x, y)\,dx\,dy\\ &= \int_A f_X(x)\,dx\end{aligned}$$

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

is thus the probability density function of X. Similarly, the probability density function of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

A variation of Proposition 2.1 states that if X and Y are random variables and g is a function of two variables, then

$$\begin{aligned} E[g(X, Y)] &= \sum_y \sum_x g(x, y) p(x, y) && \text{in the discrete case} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy && \text{in the continuous case} \end{aligned}$$

For example, if $g(X, Y) = X + Y$, then, in the continuous case,

$$\begin{aligned} E[X + Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y)\, dx\, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y)\, dx\, dy \\ &= E[X] + E[Y] \end{aligned}$$

where the first integral is evaluated by using the variation of Proposition 2.1 with $g(x, y) = x$, and the second with $g(x, y) = y$.

The same result holds in the discrete case and, combined with the corollary in Section 2.4.3, yields that for any constants a, b

$$E[aX + bY] = aE[X] + bE[Y] \tag{2.10}$$

Joint probability distributions may also be defined for n random variables. The details are exactly the same as when $n = 2$ and are left as an exercise. The corresponding result to Equation (2.10) states that if $X_1, X_2, \ldots, X_n$ are n random variables, then for any n constants $a_1, a_2, \ldots, a_n$,

$$E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E[X_1] + a_2E[X_2] + \cdots + a_nE[X_n] \tag{2.11}$$

Example 2.29 Calculate the expected sum obtained when three fair dice are rolled.

Solution: Let X denote the sum obtained. Then $X = X_1 + X_2 + X_3$ where X_i represents the value of the ith die. Thus,

$$E[X] = E[X_1] + E[X_2] + E[X_3] = 3\left(\tfrac{7}{2}\right) = \tfrac{21}{2} \quad \blacksquare$$

Example 2.30 As another example of the usefulness of Equation (2.11), let us use it to obtain the expectation of a binomial random variable having parameters n and p. Recalling that such a random variable X represents the number of successes in n trials when each trial has probability p of being a success, we have that

$$X = X_1 + X_2 + \cdots + X_n$$

where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{if the } i\text{th trial is a failure} \end{cases}$$

Hence, X_i is a Bernoulli random variable having expectation $E[X_i] = 1(p) + 0(1-p) = p$. Thus,

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = np$$

This derivation should be compared with the one presented in Example 2.17. $\blacksquare$

Example 2.31 At a party N men throw their hats into the center of a room. The hats are mixed up and each man randomly selects one. Find the expected number of men who select their own hats.

Solution: Letting X denote the number of men that select their own hats, we can best compute $E[X]$ by noting that

$$X = X_1 + X_2 + \cdots + X_N$$

where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th man selects his own hat} \\ 0, & \text{otherwise} \end{cases}$$

Now, because the ith man is equally likely to select any of the N hats, it follows that

$$P\{X_i = 1\} = P\{i\text{th man selects his own hat}\} = \frac{1}{N}$$

and so

$$E[X_i] = 1P\{X_i = 1\} + 0P\{X_i = 0\} = \frac{1}{N}$$

Hence, from Equation (2.11) we obtain that

$$E[X] = E[X_1] + \cdots + E[X_N] = \left(\frac{1}{N}\right)N = 1$$

Hence, no matter how many people are at the party, on the average exactly one of the men will select his own hat. ■

Example 2.32 Suppose there are 25 different types of coupons and suppose that each time one obtains a coupon, it is equally likely to be any one of the 25 types. Compute the expected number of different types that are contained in a set of 10 coupons.

Solution: Let X denote the number of different types in the set of 10 coupons. We compute $E[X]$ by using the representation

$$X = X_1 + \cdots + X_{25}$$

where

$$X_i = \begin{cases} 1, & \text{if at least one type } i \text{ coupon is in the set of 10} \\ 0, & \text{otherwise} \end{cases}$$

Now,

$$\begin{aligned} E[X_i] &= P\{X_i = 1\} \\ &= P\{\text{at least one type } i \text{ coupon is in the set of 10}\} \\ &= 1 - P\{\text{no type } i \text{ coupons are in the set of 10}\} \\ &= 1 - \left(\tfrac{24}{25}\right)^{10} \end{aligned}$$

when the last equality follows since each of the 10 coupons will (independently) not be a type i with probability $\frac{24}{25}$. Hence,

$$E[X] = E[X_1] + \cdots + E[X_{25}] = 25\left[1 - \left(\tfrac{24}{25}\right)^{10}\right] \quad ■$$

2.5.2. Independent Random Variables

The random variables X and Y are said to be *independent* if, for all a, b,

$$P\{X \leqslant a,\ Y \leqslant b\} = P\{X \leqslant a\}P\{Y \leqslant b\} \tag{2.12}$$

In other words, X and Y are independent if, for all a and b, the events $E_a = \{X \leqslant a\}$ and $F_b = \{Y \leqslant b\}$ are independent.

In terms of the joint distribution function F of X and Y, we have that X and Y are independent if

$$F(a,\ b) = F_X(a)F_Y(b) \qquad \text{for all } a,\ b$$

When X and Y are discrete, the condition of independence reduces to

$$p(x, y) = p_X(x)p_Y(y) \tag{2.13}$$

while if X and Y are jointly continuous, independence reduces to

$$f(x,\ y) = f_X(x)f_Y(y) \tag{2.14}$$

To prove this statement, consider first the discrete version, and suppose that the joint probability mass function $p(x,\ y)$ satisfies Equation (2.13). Then

$$\begin{aligned}
P\{X \leqslant a,\ Y \leqslant b\} &= \sum_{y \leqslant b} \sum_{x \leqslant a} p(x,\ y) \\
&= \sum_{y \leqslant b} \sum_{x \leqslant a} p_X(x)p_Y(y) \\
&= \sum_{y \leqslant b} p_Y(y) \sum_{x \leqslant a} p_X(x) \\
&= P\{Y \leqslant b\}P\{X \leqslant a\}
\end{aligned}$$

and so X and Y are independent. That Equation (2.14) implies independence in the continuous case is proven in the same manner and is left as an exercise.

An important result concerning independence is the following.

Proposition 2.3 If X and Y are independent, then for any functions h and g

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof Suppose that X and Y are jointly continuous. Then

$$\begin{aligned}
E[g(X)h(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f(x,\ y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} h(y)f_Y(y)\,dy \int_{-\infty}^{\infty} g(x)f_X(x)\,dx \\
&= E[h(Y)]E[g(X)]
\end{aligned}$$

The proof in the discrete case is similar. ■

2.5.3. Covariance and Variance of Sums of Random Variables

The covariance of any two random variables X and Y, denoted by $\text{Cov}(X, Y)$, is defined by

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])]\\ &= E[XY - YE[X] - XE[Y] + E[X]E[Y]]\\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y]\\ &= E[XY] - E[X]E[Y]\end{aligned}$$

Note that if X and Y are independent, then by Proposition 2.3 it follows that $\text{Cov}(X, Y) = 0$.

Let us consider now the special case where X and Y are indicator variables for whether or not the events A and B occur. That is, for events A and B, define

$$X = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise,} \end{cases} \qquad Y = \begin{cases} 1, & \text{if } B \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

and, because XY will equal 1 or 0 depending on whether or not both X and Y equal 1, we see that

$$\text{Cov}(X, Y) = P\{X = 1, Y = 1\} - P\{X = 1\}P\{Y = 1\}$$

From this we see that

$$\begin{aligned}\text{Cov}(X, Y) > 0 &\Leftrightarrow P\{X = 1, Y = 1\} > P\{X = 1\}P\{Y = 1\}\\ &\Leftrightarrow \frac{P\{X = 1, Y = 1\}}{P\{X = 1\}} > P\{Y = 1\}\\ &\Leftrightarrow P\{Y = 1|X = 1\} > P\{Y = 1\}\end{aligned}$$

That is, the covariance of X and Y is positive if the outcome $X = 1$ makes it more likely that $Y = 1$ (which, as is easily seen by symmetry, also implies the reverse).

In general it can be shown that a positive value of $\text{Cov}(X, Y)$ is an indication that Y tends to increase as X does, whereas a negative value indicates that Y tends to decrease as X increases.

The following are important properties of covariance.

Properties of Covariance

For any random variables X, Y, Z and constant c,

1. $\text{Cov}(X, X) = \text{Var}(X)$,
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$,
3. $\text{Cov}(cX, Y) = c\,\text{Cov}(X, Y)$,
4. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$.

Whereas the first three properties are immediate, the final one is easily proven as follows:

$$\begin{aligned}\text{Cov}(X, Y + Z) &= E[X(Y + Z)] - E[X]E[Y + Z] \\ &= E[XY] - E[X]E[Y] + E[XZ] - E[X]E[Z] \\ &= \text{Cov}(X, Y) + \text{Cov}(X, Z)\end{aligned}$$

The fourth property listed easily generalizes to give the following result:

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j) \tag{2.15}$$

A useful expression for the variance of the sum of random variables can be obtained from Equation (2.15) as follows:

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j<i} \text{Cov}(X_i, X_j) \qquad (2.16)\end{aligned}$$

If $X_i, i = 1, \ldots, n$ are independent random variables, then Equation (2.16) reduces to

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

Definition 2.1 If $X_1, \dots, X_n$ are independent and identically distributed, then the random variable $\bar{X} = \sum_{i=1}^{n} X_i/n$ is called the *sample mean*.

The following proposition shows that the covariance between the sample mean and a deviation from that sample mean is zero. It will be needed in Section 2.6.1.

Proposition 2.4 Suppose that $X_1, \dots, X_n$ are independent and identically distributed with expected value μ and variance σ^2. Then,

(a) $E[\bar{X}] = \mu$.
(b) $\mathrm{Var}(\bar{X}) = \sigma^2/n$.
(c) $\mathrm{Cov}(\bar{X}, X_i - \bar{X}) = 0, \; i = 1, \dots, n$.

Proof Parts (a) and (b) are easily established as follows:

$$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{m} E[X_i] = \mu,$$

$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

To establish part (c) we reason as follows:

$$\begin{aligned}
\mathrm{Cov}(\bar{X}, X_i - \bar{X}) &= \mathrm{Cov}(\bar{X}, X_i) - \mathrm{Cov}(\bar{X}, \bar{X}) \\
&= \frac{1}{n}\mathrm{Cov}\left(X_i + \sum_{j \neq i} X_j, X_i\right) - \mathrm{Var}(\bar{X}) \\
&= \frac{1}{n}\mathrm{Cov}(X_i, X_i) + \frac{1}{n}\mathrm{Cov}\left(\sum_{j \neq i} X_j, X_i\right) - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0
\end{aligned}$$

where the final equality used the fact that X_i and $\sum_{j \neq i} X_j$ are independent and thus have covariance 0. ■

Equation (2.16) is often useful when computing variances.

Example 2.33 (Variance of a Binomial Random Variable) Compute the variance of a binomial random variable X with parameters n and p.

Solution: Since such a random variable represents the number of successes in n independent trials when each trial has a common probability p of being a success, we may write

$$X = X_1 + \cdots + X_n$$

where the X_i are independent Bernoulli random variables such that

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{otherwise} \end{cases}$$

Hence, from Equation (2.16) we obtain

$$\text{Var}(X) = \text{Var}(X_1) + \cdots + \text{Var}(X_n)$$

But

$$\begin{aligned} \text{Var}(X_i) &= E[X_i^2] - (E[X_i])^2 \\ &= E[X_i] - (E[X_i])^2 \qquad \text{since } X_i^2 = X_i \\ &= p - p^2 \end{aligned}$$

and thus

$$\text{Var}(X) = np(1-p) \quad \blacksquare$$

Example 2.34 (Sampling from a Finite Population: The Hypergeometric) Consider a population of N individuals, some of whom are in favor of a certain proposition. In particular suppose that Np of them are in favor and $N - Np$ are opposed, where p is assumed to be unknown. We are interested in estimating p, the fraction of the population that is for the proposition, by randomly choosing and then determining the positions of n members of the population.

In such situations as described in the preceding, it is common to use the fraction of the sampled population that is in favor of the proposition as an estimator of p. Hence, if we let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th person chosen is in favor} \\ 0, & \text{otherwise} \end{cases}$$

then the usual estimator of p is $\sum_{i=1}^{n} X_i/n$. Let us now compute its mean and variance. Now

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{1}^{n} E[X_i]$$

$$= np$$

where the final equality follows since the ith person chosen is equally likely to be any of the N individuals in the population and so has probability Np/N of being in favor.

$$\mathrm{Var}\left(\sum_1^n X_i\right) = \sum_1^n \mathrm{Var}(X_i) + 2\sum_{i<j}\sum \mathrm{Cov}(X_i, X_j)$$

Now, since X_i is a Bernoulli random variable with mean p, it follows that

$$\mathrm{Var}(X_i) = p(1-p)$$

Also, for $i \neq j$,

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] \\ &= P\{X_i = 1, X_j = 1\} - p^2 \\ &= P\{X_i = 1\}P\{X_j = 1 \mid X_i = 1\} - p^2 \\ &= \frac{Np}{N}\frac{(Np-1)}{N-1} - p^2\end{aligned}$$

where the last equality follows since if the ith person to be chosen is in favor, then the jth person chosen is equally likely to be any of the other $N-1$ of which $Np-1$ are in favor. Thus, we see that

$$\begin{aligned}\mathrm{Var}\left(\sum_1^n X_i\right) &= np(1-p) + 2\binom{n}{2}\left[\frac{p(Np-1)}{N-1} - p^2\right] \\ &= np(1-p) - \frac{n(n-1)p(1-p)}{N-1}\end{aligned}$$

and so the mean and variance of our estimator are given by

$$E\left[\sum_1^n \frac{X_i}{n}\right] = p,$$

$$\mathrm{Var}\left[\sum_1^n \frac{X_i}{n}\right] = \frac{p(1-p)}{n} - \frac{(n-1)p(1-p)}{n(N-1)}$$

Some remarks are in order: As the mean of the estimator is the unknown value p, we would like its variance to be as small as possible (why is this?), and we see by the preceding that, as a function of the population size N, the variance increases as N increases. The limiting value, as $N \to \infty$, of the variance is $p(1-p)/n$, which is not surprising since for N large each of the X_i will be (approximately) independent random variables, and thus $\sum_1^n X_i$ will have an (approximately) binomial distribution with parameters n and p.

The random variable $\sum_1^n X_i$ can be thought of as representing the number of white balls obtained when n balls are randomly selected from a population consisting of Np white and $N - Np$ black balls. (Identify a person who favors the proposition with a white ball and one against with a black ball.) Such a random variable is called *hypergeometric* and has a probability mass function given by

$$P\left\{\sum_1^n X_i = k\right\} = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}} \quad \blacksquare$$

It is often important to be able to calculate the distribution of $X + Y$ from the distributions of X and Y when X and Y are independent. Suppose first that X and Y are continuous, X having probability density f and Y having probability density g. Then, letting $F_{X+Y}(a)$ be the cumulative distribution function of $X + Y$, we have

$$\begin{aligned} F_{X+Y}(a) &= P\{X + Y \leqslant a\} \\ &= \iint_{x+y\leqslant a} f(x)g(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f(x)g(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-y} f(x)\,dx\right)g(y)\,dy \\ &= \int_{-\infty}^{\infty} F_X(a-y)g(y)\,dy \end{aligned} \tag{2.17}$$

The cumulative distribution function F_{X+Y} is called the *convolution* of the distributions F_X and F_Y (the cumulative distribution functions of X and Y, respectively).

By differentiating Equation (2.17), we obtain that the probability density function $f_{X+Y}(a)$ of $X + Y$ is given by

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da}\int_{-\infty}^{\infty} F_X(a-y)g(y)\,dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da}(F_X(a-y))g(y)\,dy \\ &= \int_{-\infty}^{\infty} f(a-y)g(y)\,dy \end{aligned} \tag{2.18}$$

Example 2.35 (Sum of Two Independent Uniform Random Variables) If X and Y are independent random variables both uniformly distributed on (0, 1), then calculate the probability density of $X + Y$.

Solution: From Equation (2.18), since

$$f(a) = g(a) = \begin{cases} 1, & 0 < a < 1 \\ 0, & \text{otherwise} \end{cases}$$

we obtain

$$f_{X+Y}(a) = \int_0^1 f(a-y)\,dy$$

For $0 \leqslant a \leqslant 1$, this yields

$$f_{X+Y}(a) = \int_0^a dy = a$$

For $1 < a < 2$, we get

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2 - a$$

Hence,

$$f_{X+Y}(a) = \begin{cases} a, & 0 \leqslant a \leqslant 1 \\ 2-a, & 1 < a < 2 \\ 0, & \text{otherwise} \end{cases} \quad \blacksquare$$

Rather than deriving a general expression for the distribution of $X + Y$ in the discrete case, we shall consider an example.

Example 2.36 (Sums of Independent Poisson Random Variables) Let X and Y be independent Poisson random variables with respective means λ_1 and λ_2. Calculate the distribution of $X + Y$.

Solution: Since the event $\{X + Y = n\}$ may be written as the union of the disjoint events $\{X = k, Y = n - k\}$, $0 \leqslant k \leqslant n$, we have

$$\begin{aligned} P\{X + Y = n\} &= \sum_{k=0}^{n} P\{X = k,\ Y = n - k\} \\ &= \sum_{k=0}^{n} P\{X = k\}P\{Y = n - k\} \end{aligned}$$

$$= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!}$$
$$= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!}$$
$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}$$
$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1+\lambda_2)^n$$

In words, $X_1 + X_2$ has a Poisson distribution with mean $\lambda_1 + \lambda_2$. ■

The concept of independence may, of course, be defined for more than two random variables. In general, the n random variables $X_1, X_2, \ldots, X_n$ are said to be independent if, for all values $a_1, a_2, \ldots, a_n$,

$$P\{X_1 \leqslant a_1, X_2 \leqslant a_2, \ldots, X_n \leqslant a_n\} = P\{X_1 \leqslant a_1\} P\{X_2 \leqslant a_2\} \cdots P\{X_n \leqslant a_n\}$$

Example 2.37 Let $X_1, \ldots, X_n$ be independent and identically distributed continuous random variables with probability distribution F and density function $F' = f$. If we let $X_{(i)}$ denote the ith smallest of these random variables, then $X_{(1)}, \ldots, X_{(n)}$ are called the *order statistics*. To obtain the distribution of $X_{(i)}$, note that $X_{(i)}$ will be less than or equal to x if and only if at least i of the n random variables $X_1, \ldots, X_n$ are less than or equal to x. Hence,

$$P\{X_{(i)} \leqslant x\} = \sum_{k=i}^{n} \binom{n}{k} (F(x))^k (1 - F(x))^{n-k}$$

Differentiation yields that the density function of $X_{(i)}$ is as follows:

$$f_{X_{(i)}}(x) = f(x) \sum_{k=i}^{n} \binom{n}{k} k (F(x))^{k-1} (1 - F(x))^{n-k}$$
$$- f(x) \sum_{k=i}^{n} \binom{n}{k} (n-k) (F(x))^k (1 - F(x))^{n-k-1}$$
$$= f(x) \sum_{k=i}^{n} \frac{n!}{(n-k)!(k-1)!} (F(x))^{k-1} (1 - F(x))^{n-k}$$
$$- f(x) \sum_{k=i}^{n-1} \frac{n!}{(n-k-1)!k!} (F(x))^k (1 - F(x))^{n-k-1}$$

$$= f(x) \sum_{k=i}^{n} \frac{n!}{(n-k)!(k-1)!} (F(x))^{k-1} (1 - F(x))^{n-k}$$

$$- f(x) \sum_{j=i+1}^{n} \frac{n!}{(n-j)!(j-1)!} (F(x))^{j-1} (1 - F(x))^{n-j}$$

$$= \frac{n!}{(n-i)!(i-1)!} f(x)(F(x))^{i-1} (1 - F(x))^{n-i}$$

The preceding density is quite intuitive, since in order for $X_{(i)}$ to equal x, $i-1$ of the n values $X_1, \ldots, X_n$ must be less than x; $n-i$ of them must be greater than x; and one must be equal to x. Now, the probability density that every member of a specified set of $i-1$ of the X_j is less than x, every member of another specified set of $n-i$ is greater than x, and the remaining value is equal to x is $(F(x))^{i-1}(1 - F(x))^{n-i} f(x)$. Therefore, since there are $n!/[(i-1)!(n-i)!]$ different partitions of the n random variables into the three groups, we obtain the preceding density function. ■

2.5.4. Joint Probability Distribution of Functions of Random Variables

Let X_1 and X_2 be jointly continuous random variables with joint probability density function $f(x_1, x_2)$. It is sometimes necessary to obtain the joint distribution of the random variables Y_1 and Y_2 which arise as functions of X_1 and X_2. Specifically, suppose that $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ for some functions g_1 and g_2.

Assume that the functions g_1 and g_2 satisfy the following conditions:

1. The equations $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ can be uniquely solved for x_1 and x_2 in terms of y_1 and y_2 with solutions given by, say, $x_1 = h_1(y_1, y_2)$, $x_2 = h_2(y_1, y_2)$.
2. The functions g_1 and g_2 have continuous partial derivatives at all points (x_1, x_2) and are such that the following 2×2 determinant

$$J(x_1, x_2) = \begin{vmatrix} \dfrac{\partial g_1}{\partial x_1} & \dfrac{\partial g_1}{\partial x_2} \\ \dfrac{\partial g_2}{\partial x_1} & \dfrac{\partial g_2}{\partial x_2} \end{vmatrix} \equiv \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1} \neq 0$$

at all points (x_1, x_2).

Under these two conditions it can be shown that the random variables Y_1 and Y_2 are jointly continuous with joint density function given by

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|J(x_1, x_2)|^{-1} \tag{2.19}$$

where $x_1 = h_1(y_1, y_2), x_2 = h_2(y_1, y_2)$.

A proof of Equation (2.19) would proceed along the following lines:

$$P\{Y_1 \leqslant y_1,\ Y_2 \leqslant y_2\} = \iint\limits_{\substack{(x_1,x_2):\\ g_1(x_1,x_2)\leqslant y_1\\ g_2(x_1,x_2)\leqslant y_2}} f_{X_1,\,X_2}(x_1,\ x_2)\,dx_1\,dx_2 \tag{2.20}$$

The joint density function can now be obtained by differentiating Equation (2.20) with respect to y_1 and y_2. That the result of this differentiation will be equal to the right-hand side of Equation (2.19) is an exercise in advanced calculus whose proof will not be presented in the present text.

Example 2.38 If X and Y are independent gamma random variables with parameters (α, λ) and (β, λ), respectively, compute the joint density of $U = X + Y$ and $V = X/(X + Y)$.

Solution: The joint density of X and Y is given by

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} \frac{\lambda e^{-\lambda y}(\lambda y)^{\beta-1}}{\Gamma(\beta)} \\ &= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda(x+y)} x^{\alpha-1} y^{\beta-1} \end{aligned}$$

Now, if $g_1(x, y) = x + y,\ g_2(x, y) = x/(x + y)$, then

$$\frac{\partial g_1}{\partial x} = \frac{\partial g_1}{\partial y} = 1, \qquad \frac{\partial g_2}{\partial x} = \frac{y}{(x+y)^2}, \qquad \frac{\partial g_2}{\partial y} = -\frac{x}{(x+y)^2}$$

and so

$$J(x,\ y) = \begin{vmatrix} 1 & 1 \\ \dfrac{y}{(x+y)^2} & \dfrac{-x}{(x+y)^2} \end{vmatrix} = -\frac{1}{x+y}$$

Finally, because the equations $u = x + y,\ v = x/(x + y)$ have as their solutions $x = uv,\ y = u(1 - v)$, we see that

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}[uv,\ u(1 - v)]u \\ &= \frac{\lambda e^{-\lambda u}(\lambda u)^{\alpha+\beta-1}}{\Gamma(\alpha + \beta)} \frac{v^{\alpha-1}(1 - v)^{\beta-1}\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \end{aligned}$$

Hence $X + Y$ and $X/(X + Y)$ are independent, with $X + Y$ having a gamma distribution with parameters $(\alpha + \beta, \lambda)$ and $X/(X + Y)$ having density function

$$f_V(v) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha-1}(1 - v)^{\beta-1}, \qquad 0 < v < 1$$

This is called the beta density with parameters (α, β).

This result is quite interesting. For suppose there are $n + m$ jobs to be performed, with each (independently) taking an exponential amount of time with rate λ for performance, and suppose that we have two workers to perform these jobs. Worker I will do jobs $1, 2, \ldots, n$, and worker II will do the remaining m jobs. If we let X and Y denote the total working times of workers I and II, respectively, then upon using the preceding result it follows that X and Y will be independent gamma random variables having parameters (n, λ) and (m, λ), respectively. Then the preceding result yields that independently of the working time needed to complete all $n + m$ jobs (that is, of $X + Y$), the proportion of this work that will be performed by worker I has a beta distribution with parameters (n, m). ■

When the joint density function of the n random variables $X_1, X_2, \ldots, X_n$ is given and we want to compute the joint density function of $Y_1, Y_2, \ldots, Y_n$, where

$$Y_1 = g_1(X_1, \ldots, X_n), \qquad Y_2 = g_2(X_1, \ldots, X_n), \qquad \ldots,$$
$$Y_n = g_n(X_1, \ldots, X_n)$$

the approach is the same. Namely, we assume that the functions g_i have continuous partial derivatives and that the Jacobian determinant $J(x_1, \ldots, x_n) \neq 0$ at all points $(x_1, \ldots, x_n)$, where

$$J(x_1, \ldots, x_n) = \begin{vmatrix} \dfrac{\partial g_1}{\partial x_1} & \dfrac{\partial g_1}{\partial x_2} & \cdots & \dfrac{\partial g_1}{\partial x_n} \\ \dfrac{\partial g_2}{\partial x_1} & \dfrac{\partial g_2}{\partial x_2} & \cdots & \dfrac{\partial g_2}{\partial x_n} \\ \dfrac{\partial g_n}{\partial x_1} & \dfrac{\partial g_n}{\partial x_2} & \cdots & \dfrac{\partial g_n}{\partial x_n} \end{vmatrix}$$

Furthermore, we suppose that the equations $y_1 = g_1(x_1, \ldots, x_n)$, $y_2 = g_2(x_1, \ldots, x_n), \ldots, y_n = g_n(x_1, \ldots, x_n)$ have a unique solution, say, $x_1 = h_1(y_1, \ldots, y_n), \ldots, x_n = h_n(y_1, \ldots, y_n)$. Under these assumptions the joint density function of the random variables Y_i is given by

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\,|J(x_1, \ldots, x_n)|^{-1}$$

where $x_i = h_i(y_1, \ldots, y_n)$, $i = 1, 2, \ldots, n$.

2.6. Moment Generating Functions

The moment generating function $\phi(t)$ of the random variable X is defined for all values t by

$$\begin{aligned}
\phi(t) &= E[e^{tX}] \\
&= \begin{cases} \displaystyle\sum_x e^{tx} p(x), & \text{if } X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} e^{tx} f(x)\,dx, & \text{if } X \text{ is continuous} \end{cases}
\end{aligned}$$

We call $\phi(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $\phi(t)$. For example,

$$\begin{aligned}
\phi'(t) &= \frac{d}{dt} E[e^{tX}] \\
&= E\left[\frac{d}{dt}(e^{tX})\right] \\
&= E[Xe^{tX}]
\end{aligned}$$

Hence,

$$\phi'(0) = E[X]$$

Similarly,

$$\begin{aligned}
\phi''(t) &= \frac{d}{dt}\phi'(t) \\
&= \frac{d}{dt} E[Xe^{tX}] \\
&= E\left[\frac{d}{dt}(Xe^{tX})\right] \\
&= E[X^2 e^{tX}]
\end{aligned}$$

and so

$$\phi''(0) = E[X^2]$$

In general, the nth derivative of $\phi(t)$ evaluated at $t = 0$ equals $E[X^n]$, that is,

$$\phi^n(0) = E[X^n], \qquad n \geqslant 1$$

We now compute $\phi(t)$ for some common distributions.

Example 2.39 (The Binomial Distribution with Parameters n and p)

$$\begin{aligned}
\phi(t) &= E[e^{tX}] \\
&= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\
&= (pe^t + 1 - p)^n
\end{aligned}$$

Hence,

$$\phi'(t) = n(pe^t + 1 - p)^{n-1} pe^t$$

and so

$$E[X] = \phi'(0) = np$$

which checks with the result obtained in Example 2.17. Differentiating a second time yields

$$\phi''(t) = n(n-1)(pe^t + 1 - p)^{n-2}(pe^t)^2 + n(pe^t + 1 - p)^{n-1} pe^t$$

and so

$$E[X^2] = \phi''(0) = n(n-1)p^2 + np$$

Thus, the variance of X is given

$$\begin{aligned}
\mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\
&= n(n-1)p^2 + np - n^2p^2 \\
&= np(1-p) \quad \blacksquare
\end{aligned}$$

Example 2.40 (The Poisson Distribution with Mean λ)

$$\begin{aligned}
\phi(t) &= E[e^{tX}] \\
&= \sum_{n=0}^{\infty} \frac{e^{tn} e^{-\lambda} \lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\
&= e^{-\lambda} e^{\lambda e^t} \\
&= \exp\{\lambda(e^t - 1)\}
\end{aligned}$$

Differentiation yields

$$\begin{aligned}
\phi'(t) &= \lambda e^t \exp\{\lambda(e^t - 1)\}, \\
\phi''(t) &= (\lambda e^t)^2 \exp\{\lambda(e^t - 1)\} + \lambda e^t \exp\{\lambda(e^t - 1)\}
\end{aligned}$$

and so

$$\begin{aligned}
E[X] &= \phi'(0) = \lambda, \\
E[X^2] &= \phi''(0) = \lambda^2 + \lambda, \\
\operatorname{Var}(X) &= E[X^2] - (E[X])^2 \\
&= \lambda
\end{aligned}$$

Thus, both the mean and the variance of the Poisson equal λ. ■

Example 2.41 (The Exponential Distribution with Parameter λ)

$$\begin{aligned}
\phi(t) &= E[e^{tX}] \\
&= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} \, dx \\
&= \lambda \int_0^{\infty} e^{-(\lambda - t)x} \, dx \\
&= \frac{\lambda}{\lambda - t} \qquad \text{for } t < \lambda
\end{aligned}$$

We note by the preceding derivation that, for the exponential distribution, $\phi(t)$ is only defined for values of t less than λ. Differentiation of $\phi(t)$ yields

$$\phi'(t) = \frac{\lambda}{(\lambda - t)^2}, \qquad \phi''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

Hence,

$$E[X] = \phi'(0) = \frac{1}{\lambda}, \qquad E[X^2] = \phi''(0) = \frac{2}{\lambda^2}$$

The variance of X is thus given by

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \frac{1}{\lambda^2} \quad \blacksquare$$

Example 2.42 (The Normal Distribution with Parameters μ and σ^2) The moment generating function of a standard normal random variable Z is obtained as follows.

$$\begin{aligned} E[e^{tZ}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2-2tx)/2}\, dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2}\, dx \\ &= e^{t^2/2} \end{aligned}$$

If Z is a standard normal, then $X = \sigma Z + \mu$ is normal with parameters μ and σ^2; therefore,

$$\phi(t) = E[e^{tX}] = E[e^{t(\sigma Z + \mu)}] = e^{t\mu} E[e^{t\sigma Z}] = \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\}$$

By differentiating we obtain

$$\begin{aligned} \phi'(t) &= (\mu + t\sigma^2) \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\}, \\ \phi''(t) &= (\mu + t\sigma^2)^2 \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\} + \sigma^2 \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\} \end{aligned}$$

and so

$$E[X] = \phi'(0) = \mu,$$
$$E[X^2] = \phi''(0) = \mu^2 + \sigma^2$$

implying that

$$\begin{aligned}\mathrm{Var}(X) &= E[X^2] - E([X])^2 \\ &= \sigma^2 \quad \blacksquare\end{aligned}$$

Tables 2.1 and 2.2 give the moment generating function for some common distributions.

An important property of moment generating functions is that the *moment generating function of the sum of independent random variables is just the product of the individual moment generating functions*. To see this, suppose that X and Y are independent and have moment generating functions $\phi_X(t)$ and $\phi_Y(t)$, respectively. Then $\phi_{X+Y}(t)$, the moment generating function of $X + Y$, is given by

$$\begin{aligned}\phi_{X+Y}(t) &= E[e^{t(X+Y)}] \\ &= E[e^{tX}e^{tY}] \\ &= E[e^{tX}]E[e^{tY}] \\ &= \phi_X(t)\phi_Y(t)\end{aligned}$$

where the next to the last equality follows from Proposition 2.3 since X and Y are independent.

Table 2.1

Discrete probability distribution	Probability mass function, $p(x)$	Moment generating function, $\phi(t)$	Mean	Variance
Binomial with parameters n, p $0 \leqslant p \leqslant 1$	$\binom{n}{x}p^x(1-p)^{n-x}$, $x = 0, 1, \dots, n$	$(pe^t + (1-p))^n$	np	$np(1-p)$
Poisson with parameter $\lambda > 0$	$e^{-\lambda}\dfrac{\lambda^x}{x!}$, $x = 0, 1, 2, \dots$	$\exp\{\lambda(e^t - 1)\}$	λ	λ
Geometric with parameter $0 \leqslant p \leqslant 1$	$p(1-p)^{x-1}$, $x = 1, 2, \dots$	$\dfrac{pe^t}{1-(1-p)e^t}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$

Table 2.2

Continuous probability distribution	Probability density function, $f(x)$	Moment generating function, $\phi(t)$	Mean	Variance
Uniform over (a, b)	$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$	$\dfrac{e^{tb} - e^{ta}}{t(b-a)}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x < 0 \end{cases}$	$\dfrac{\lambda}{\lambda - t}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Gamma with parameters (n, λ) $\lambda > 0$	$f(x) = \begin{cases} \dfrac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$	$\left(\dfrac{\lambda}{\lambda - t}\right)^n$	$\dfrac{n}{\lambda}$	$\dfrac{n}{\lambda^2}$
Normal with parameters (μ, σ^2)	$f(x) = \dfrac{1}{\sqrt{2\pi}\,\sigma} \times \exp\{-(x-\mu)^2/2\sigma^2\}$, $-\infty < x < \infty$	$\exp\left\{\mu t + \dfrac{\sigma^2 t^2}{2}\right\}$	μ	σ^2

Another important result is that the *moment generating function uniquely determines the distribution*. That is, there exists a one-to-one correspondence between the moment generating function and the distribution function of a random variable.

Example 2.43 Suppose the moment generating function of a random variable X is given by $\phi(t) = e^{3(e^t - 1)}$. What is $P\{X = 0\}$?

Solution: We see from Table 2.1 that $\phi(t) = e^{3(e^t-1)}$ is the moment generating function of a Poisson random variable with mean 3. Hence, by the one-to-one correspondence between moment generating functions and distribution functions, it follows that X must be a Poisson random variable with mean 3. Thus, $P\{X = 0\} = e^{-3}$. ■

Example 2.44 (Sums of Independent Binomial Random Variables) If X and Y are independent binomial random variables with parameters (n, p) and (m, p), respectively, then what is the distribution of $X + Y$?

Solution: The moment generating function of $X + Y$ is given by

$$\begin{aligned} \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) &= (pe^t + 1 - p)^n (pe^t + 1 - p)^m \\ &= (pe^t + 1 - p)^{m+n} \end{aligned}$$

But $(pe^t + (1-p))^{m+n}$ is just the moment generating function of a binomial random variable having parameters $m+n$ and p. Thus, this must be the distribution of $X+Y$. ■

Example 2.45 (Sums of Independent Poisson Random Variables) Calculate the distribution of $X+Y$ when X and Y are independent Poisson random variables with means λ_1 and λ_2, respectively.

Solution:

$$\begin{aligned}\phi_{X+Y}(t) &= \phi_X(t)\,\phi_Y(t)\\ &= e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)}\\ &= e^{(\lambda_1+\lambda_2)(e^t-1)}\end{aligned}$$

Hence, $X+Y$ is Poisson distributed with mean $\lambda_1+\lambda_2$, verifying the result given in Example 2.36. ■

Example 2.46 (Sums of Independent Normal Random Variables) Show that if X and Y are independent normal random variables with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively, then $X+Y$ is normal with mean $\mu_1+\mu_2$ and variance $\sigma_1^2+\sigma_2^2$.

Solution:

$$\begin{aligned}\phi_{X+Y}(t) &= \phi_X(t)\phi_Y(t)\\ &= \exp\left\{\frac{\sigma_1^2t^2}{2}+\mu_1 t\right\}\exp\left\{\frac{\sigma_2^2t^2}{2}+\mu_2 t\right\}\\ &= \exp\left\{\frac{(\sigma_1^2+\sigma_2^2)t^2}{2}+(\mu_1+\mu_2)t\right\}\end{aligned}$$

which is the moment generating function of a normal random variable with mean $\mu_1+\mu_2$ and variance $\sigma_1^2+\sigma_2^2$. Hence, the result follows since the moment generating function uniquely determines the distribution. ■

Example 2.47 (The Poisson Paradigm) We showed in Section 2.2.4 that the number of successes that occur in n independent trails, each of which results in a success with probability p is, when n is large and p small, approximately a Poisson random variable with parameter $\lambda = np$. This result, however, can be substantially strengthened. First it is not necessary that the trials have the same success probability, only that all the success probabilities are small. To see that this is the case, suppose that the trials are independent, with trial i resulting in a success with probability p_i, where all the p_i, $i = 1, \dots, n$ are small. Letting X_i equal 1 if trial i is a success, and 0 otherwise, it follows that the number of successes, call it X, can be expressed as

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

Using that X_i is a Bernoulli (or binary) random variable, its moment generating function is

$$E[e^{tX_i}] = p_i e^t + 1 - p_i = 1 + p_i(e^t - 1)$$

Now, using the result that, for $|x|$ small,

$$e^x \approx 1 + x$$

it follows, because $p_i(e^t - 1)$ is small when p_i is small, that

$$E[e^{tX_i}] = 1 + p_i(e^t - 1) \approx \exp\{p_i(e^t - 1)\}$$

Because the moment generating function of a sum of independent random variables is the product of their moment generating functions, the preceding implies that

$$E[e^{tX}] \approx \prod_{i=1}^{n} \exp\{p_i(e^t - 1)\} = \exp\left\{\sum_i p_i(e^t - 1)\right\}$$

But the right side of the preceding is the moment generating function of a Poisson random variable with mean $\sum_i p_i$, thus arguing that this is approximately the distribution of X.

Not only is it not necessary for the trials to have the same success probability for the number of successes to approximately have a Poisson distribution, they need not even be independent, provided that their dependence is *weak*. For instance, recall the matching problem (Example 2.31) where n people randomly select hats from a set consisting of one hat from each person. By regarding the random selections of hats as constituting n trials, where we say that trial i is a success if person i chooses his or her own hat, it follows that, with A_i being the event that trial i is a success,

$$P(A_i) = \frac{1}{n} \quad \text{and} \quad P(A_i|A_j) = \frac{1}{n-1}, \qquad j \neq i$$

Hence, whereas the trials are not independent, their dependence appears, for large n, to be weak. Because of this weak dependence, and the small trial success probabilities, it would seem that the number of matches should approximately have a

Poisson distribution with mean 1 when n is large, and this is shown to be the case in Example 3.23.

The statement that "the number of successes in n trials that are either independent or at most weakly dependent is, when the trial success probabilities are all small, approximately a Poisson random variable" is known as the *Poisson paradigm*. ■

Remark For a nonnegative random variable X, it is often convenient to define its *Laplace transform* $g(t)$, $t \geqslant 0$, by

$$g(t) = \phi(-t) = E[e^{-tX}]$$

That is, the Laplace transform evaluated at t is just the moment generating function evaluated at $-t$. The advantage of dealing with the Laplace transform, rather than the moment generating function, when the random variable is nonnegative is that if $X \geqslant 0$ and $t \geqslant 0$, then

$$0 \leqslant e^{-tX} \leqslant 1$$

That is, the Laplace transform is always between 0 and 1. As in the case of moment generating functions, it remains true that nonnegative random variables that have the same Laplace transform must also have the same distribution. ■

It is also possible to define the joint moment generating function of two or more random variables. This is done as follows. For any n random variables $X_1, \ldots, X_n$, the joint moment generating function, $\phi(t_1, \ldots, t_n)$, is defined for all real values of $t_1, \ldots, t_n$ by

$$\phi(t_1, \ldots, t_n) = E[e^{(t_1 X_1 + \cdots + t_n X_n)}]$$

It can be shown that $\phi(t_1, \ldots, t_n)$ uniquely determines the joint distribution of $X_1, \ldots, X_n$.

Example 2.48 (The Multivariate Normal Distribution) Let $Z_1, \ldots, Z_n$ be a set of n independent standard normal random variables. If, for some constants a_{ij}, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$, and μ_i, $1 \leqslant i \leqslant m$,

$$\begin{aligned}
X_1 &= a_{11} Z_1 + \cdots + a_{1n} Z_n + \mu_1, \\
X_2 &= a_{21} Z_1 + \cdots + a_{2n} Z_n + \mu_2, \\
&\vdots \\
X_i &= a_{i1} Z_1 + \cdots + a_{in} Z_n + \mu_i, \\
&\vdots \\
X_m &= a_{m1} Z_1 + \cdots + a_{mn} Z_n + \mu_m
\end{aligned}$$

then the random variables $X_1, \ldots, X_m$ are said to have a multivariate normal distribution.

It follows from the fact that the sum of independent normal random variables is itself a normal random variable that each X_i is a normal random variable with mean and variance given by

$$E[X_i] = \mu_i,$$
$$\mathrm{Var}(X_i) = \sum_{j=1}^{n} a_{ij}^2$$

Let us now determine

$$\phi(t_1, \ldots, t_m) = E[\exp\{t_1 X_1 + \cdots + t_m X_m\}]$$

the joint moment generating function of $X_1, \ldots, X_m$. The first thing to note is that since $\sum_{i=1}^{m} t_i X_i$ is itself a linear combination of the independent normal random variables $Z_1, \ldots, Z_n$, it is also normally distributed. Its mean and variance are respectively

$$E\left[\sum_{i=1}^{m} t_i X_i\right] = \sum_{i=1}^{m} t_i \mu_i$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{m} t_i X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^{m} t_i X_i, \sum_{j=1}^{m} t_j X_j\right)$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{m} t_i t_j \mathrm{Cov}(X_i, X_j)$$

Now, if Y is a normal random variable with mean μ and variance σ^2, then

$$E[e^Y] = \phi_Y(t)|_{t=1} = e^{\mu+\sigma^2/2}$$

Thus, we see that

$$\phi(t_1, \ldots, t_m) = \exp\left\{\sum_{i=1}^{m} t_i \mu_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} t_i t_j \mathrm{Cov}(X_i, X_j)\right\}$$

which shows that the joint distribution of $X_1, \ldots, X_m$ is completely determined from a knowledge of the values of $E[X_i]$ and $\mathrm{Cov}(X_i, X_j)$, $i, j = 1, \ldots, m$. ■

2.6.1. The Joint Distribution of the Sample Mean and Sample Variance from a Normal Population

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and variance σ^2. The random variable S^2 defined by

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

is called the *sample variance* of these data. To compute $E[S^2]$ we use the identity

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \tag{2.21}$$

which is proven as follows:

$$\begin{aligned}
\sum_{i=1}^{n} (X_i - \bar{X}) &= \sum_{i=1}^{n} (X_i - \mu + \mu - \bar{X})^2 \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X}) \sum_{i=1}^{n} (X_i - \mu) \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\mu - \bar{X})^2 + 2(\mu - \bar{X})(n\bar{X} - n\mu) \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\mu - \bar{X})^2 - 2n(\mu - \bar{X})^2
\end{aligned}$$

and Identity (2.21) follows.

Using Identity (2.21) gives

$$\begin{aligned}
E[(n-1)S^2] &= \sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \\
&= n\sigma^2 - n\,\mathrm{Var}(\bar{X}) \\
&= (n-1)\sigma^2 \qquad \text{from Proposition 2.4(b)}
\end{aligned}$$

Thus, we obtain from the preceding that

$$E[S^2] = \sigma^2$$

We will now determine the joint distribution of the sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ and the sample variance S^2 when the X_i have a normal distribution. To begin we need the concept of a chi-squared random variable.

Definition 2.2 If $Z_1, \ldots, Z_n$ are independent standard normal random variables, then the random variable $\sum_{i=1}^{n} Z_i^2$ is said to be a *chi-squared random variable* with *n degrees of freedom.*

We shall now compute the moment generating function of $\sum_{i=1}^{n} Z_i^2$. To begin, note that

$$
\begin{aligned}
E[\exp\{tZ_i^2\}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-x^2/2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2}\, dx \qquad \text{where } \sigma^2 = (1-2t)^{-1} \\
&= \sigma \\
&= (1-2t)^{-1/2}
\end{aligned}
$$

Hence,

$$
E\left[\exp\left\{t\sum_{i=1}^{n} Z_i^2\right\}\right] = \prod_{i=1}^{n} E[\exp\{tZ_i^2\}] = (1-2t)^{-n/2}
$$

Now, let $X_1, \ldots, X_n$ be independent normal random variables, each with mean μ and variance σ^2, and let $\bar{X} = \sum_{i=1}^{n} X_i/n$ and S^2 denote their sample mean and sample variance. Since the sum of independent normal random variables is also a normal random variable, it follows that $\bar{X}$ is a normal random variable with expected value μ and variance σ^2/n. In addition, from Proposition 2.4,

$$
\text{Cov}(\bar{X}, X_i - \bar{X}) = 0, \qquad i = 1, \ldots, n \tag{2.22}
$$

Also, since $\bar{X}, X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$ are all linear combinations of the independent standard normal random variables $(X_i - \mu)/\sigma, i = 1, \ldots, n$, it follows that the random variables $\bar{X}, X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$ have a joint distribution that is multivariate normal. However, if we let Y be a normal random variable with mean μ and variance σ^2/n that is independent of $X_1, \ldots, X_n$, then the random variables $Y, X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$ also have a multivariate normal distribution, and by Equation (2.22), they have the same expected values and covariances as the random variables $\bar{X}, X_i - \bar{X},\ i = 1, \ldots, n$. Thus, since a multivariate normal distribution is completely determined by its expected values and covariances, we can conclude that the random vectors $Y, X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$ and $\bar{X}, X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X}$ have the same joint distribution; thus showing that $\bar{X}$ is independent of the sequence of deviations $X_i - \bar{X}, i = 1, \ldots, n$.

Since $\bar{X}$ is independent of the sequence of deviations $X_i - \bar{X}, i = 1, \ldots, n$, it follows that it is also independent of the sample variance

$$S^2 \equiv \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

To determine the distribution of S^2, use Identity (2.21) to obtain

$$(n-1)S^2 = \sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Dividing both sides of this equation by σ^2 yields

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \tag{2.23}$$

Now, $\sum_{i=1}^{n} (X_i - \mu)^2/\sigma^2$ is the sum of the squares of n independent standard normal random variables, and so is a chi-squared random variable with n degrees of freedom; it thus has moment generating function $(1-2t)^{-n/2}$. Also $[(\bar{X} - \mu)/(\sigma/\sqrt{n})]^2$ is the square of a standard normal random variable and so is a chi-squared random variable with one degree of freedom; and thus has moment generating function $(1-2t)^{-1/2}$. In addition, we have previously seen that the two random variables on the left side of Equation (2.23) are independent. Therefore, because the moment generating function of the sum of independent random variables is equal to the product of their individual moment generating functions, we obtain that

$$E[e^{t(n-1)S^2/\sigma^2}](1-2t)^{-1/2} = (1-2t)^{-n/2}$$

or

$$E[e^{t(n-1)S^2/\sigma^2}] = (1-2t)^{-(n-1)/2}$$

But because $(1-2t)^{-(n-1)/2}$ is the moment generating function of a chi-squared random variable with $n-1$ degrees of freedom, we can conclude, since the moment generating function uniquely determines the distribution of the random variable, that this is the distribution of $(n-1)S^2/\sigma^2$.

Summing up, we have shown the following.

Proposition 2.5 If $X_1, \ldots, X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then the sample mean $\bar{X}$ and the sample variance S^2 are independent. $\bar{X}$ is a normal random variable with mean μ and variance σ^2/n; $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $n-1$ degrees of freedom.

2.7. Limit Theorems

We start this section by proving a result known as Markov's inequality.

Proposition 2.6 (Markov's Inequality) If X is a random variable that takes only nonnegative values, then for any value $a > 0$

$$P\{X \geqslant a\} \leqslant \frac{E[X]}{a}$$

Proof We give a proof for the case where X is continuous with density f:

$$\begin{aligned} E[X] &= \int_0^\infty x f(x)\, dx \\ &= \int_0^a x f(x)\, dx + \int_a^\infty x f(x)\, dx \\ &\geqslant \int_a^\infty x f(x)\, dx \\ &\geqslant \int_a^\infty a f(x)\, dx \\ &= a \int_a^\infty f(x)\, dx \\ &= a P\{X \geqslant a\} \end{aligned}$$

and the result is proven. ■

As a corollary, we obtain the following.

Proposition 2.7 (Chebyshev's Inequality) If X is a random variable with mean μ and variance σ^2, then, for any value $k > 0$,

$$P\{|X - \mu| \geqslant k\} \leqslant \frac{\sigma^2}{k^2}$$

Proof Since $(X - \mu)^2$ is a nonnegative random variable, we can apply Markov's inequality (with $a = k^2$) to obtain

$$P\{(X - \mu)^2 \geqslant k^2\} \leqslant \frac{E[(X - \mu)^2]}{k^2}$$

But since $(X - \mu)^2 \geqslant k^2$ if and only if $|X - \mu| \geqslant k$, the preceding is equivalent to

$$P\{|X - \mu| \geqslant k\} \leqslant \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

and the proof is complete. ■

The importance of Markov's and Chebyshev's inequalities is that they enable us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities could be exactly computed, and we would not need to resort to bounds.

Example 2.49 Suppose we know that the number of items produced in a factory during a week is a random variable with mean 500.

(a) What can be said about the probability that this week's production will be at least 1000?
(b) If the variance of a week's production is known to equal 100, then what can be said about the probability that this week's production will be between 400 and 600?

Solution: Let X be the number of items that will be produced in a week.

(a) By Markov's inequality,

$$P\{X \geqslant 1000\} \leqslant \frac{E[X]}{1000} = \frac{500}{1000} = \frac{1}{2}$$

(b) By Chebyshev's inequality,

$$P\{|X - 500| \geqslant 100\} \leqslant \frac{\sigma^2}{(100)^2} = \frac{1}{100}$$

Hence,

$$P\{|X - 500| < 100\} \geqslant 1 - \frac{1}{100} = \frac{99}{100}$$

and so the probability that this week's production will be between 400 and 600 is at least 0.99. ■

The following theorem, known as the *strong law of large numbers*, is probably the most well-known result in probability theory. It states that the average of a sequence of independent random variables having the same distribution will, with probability 1, converge to the mean of that distribution.

Theorem 2.1 (Strong Law of Large Numbers) Let $X_1, X_2, \ldots$ be a sequence of independent random variables having a common distribution, and let $E[X_i] = \mu$. Then, with probability 1,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \qquad \text{as } n \to \infty$$

As an example of the preceding, suppose that a sequence of independent trials is performed. Let E be a fixed event and denote by $P(E)$ the probability that E occurs on any particular trial. Letting

$$X_i = \begin{cases} 1, & \text{if } E \text{ occurs on the } i\text{th trial} \\ 0, & \text{if } E \text{ does not occur on the } i\text{th trial} \end{cases}$$

we have by the strong law of large numbers that, with probability 1,

$$\frac{X_1 + \cdots + X_n}{n} \to E[X] = P(E) \tag{2.24}$$

Since $X_1 + \cdots + X_n$ represents the number of times that the event E occurs in the first n trials, we may interpret Equation (2.24) as stating that, with probability 1, the limiting proportion of time that the event E occurs is just $P(E)$.

Running neck and neck with the strong law of large numbers for the honor of being probability theory's number one result is the *central limit theorem*. Besides its theoretical interest and importance, this theorem provides a simple method for computing approximate probabilities for sums of independent random variables. It also explains the remarkable fact that the empirical frequencies of so many natural "populations" exhibit a bell-shaped (that is, normal) curve.

Theorem 2.2 (Central Limit Theorem) Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables, each with mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is,

$$P\left\{\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\, dx$$

as $n \to \infty$.

Note that like the other results of this section, this theorem holds for *any* distribution of the X_i's; herein lies its power.

If X is binomially distributed with parameters n and p, then X has the same distribution as the sum of n independent Bernoulli random variables, each with parameter p. (Recall that the Bernoulli random variable is just a binomial random variable whose parameter n equals 1.) Hence, the distribution of

$$\frac{X - E[X]}{\sqrt{\text{Var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches the standard normal distribution as n approaches ∞. The normal approximation will, in general, be quite good for values of n satisfying $np(1-p) \geqslant 10$.

Example 2.50 (Normal Approximation to the Binomial) Let X be the number of times that a fair coin, flipped 40 times, lands heads. Find the probability that $X = 20$. Use the normal approximation and then compare it to the exact solution.

Solution: Since the binomial is a discrete random variable, and the normal a continuous random variable, it leads to a better approximation to write the desired probability as

$$\begin{aligned} P\{X = 20\} &= P\{19.5 < X < 20.5\} \\ &= P\left\{\frac{19.5 - 20}{\sqrt{10}} < \frac{X - 20}{\sqrt{10}} < \frac{20.5 - 20}{\sqrt{10}}\right\} \\ &= P\left\{-0.16 < \frac{X - 20}{\sqrt{10}} < 0.16\right\} \\ &\approx \Phi(0.16) - \Phi(-0.16) \end{aligned}$$

where $\Phi(x)$, the probability that the standard normal is less than x is given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy$$

By the symmetry of the standard normal distribution

$$\Phi(-0.16) = P\{N(0,\ 1) > 0.16\} = 1 - \Phi(0.16)$$

where $N(0,\ 1)$ is a standard normal random variable. Hence, the desired probability is approximated by

$$P\{X = 20\} \approx 2\Phi(0.16) - 1$$

Using Table 2.3, we obtain that

$$P\{X = 20\} \approx 0.1272$$

Table 2.3 Area $\Phi(x)$ under the Standard Normal Curve to the Left of x

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5597	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8557	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

The exact result is

$$P\{X = 20\} = \binom{40}{20}\left(\frac{1}{2}\right)^{40}$$

which can be shown to equal 0.1268. ■

Example 2.51 Let $X_i, i = 1, 2, \ldots, 10$ be independent random variables, each being uniformly distributed over (0, 1). Estimate $P\{\sum_1^{10} X_i > 7\}$.

Solution: Since $E[X_i] = \frac{1}{2}$, $\text{Var}(X_i) = \frac{1}{12}$ we have by the central limit theorem that

$$P\left\{\sum_1^{10} X_i > 7\right\} = P\left\{\frac{\sum_1^{10} X_i - 5}{\sqrt{10\left(\frac{1}{12}\right)}} > \frac{7-5}{\sqrt{10\left(\frac{1}{12}\right)}}\right\}$$
$$\approx 1 - \Phi(2.2)$$
$$= 0.0139 \quad \blacksquare$$

Example 2.52 The lifetime of a special type of battery is a random variable with mean 40 hours and standard deviation 20 hours. A battery is used until it fails, at which point it is replaced by a new one. Assuming a stockpile of 25 such batteries, the lifetimes of which are independent, approximate the probability that over 1100 hours of use can be obtained.

Solution: If we let X_i denote the lifetime of the ith battery to be put in use, then we desire $p = P\{X_1 + \cdots + X_{25} > 1100\}$, which is approximated as follows:

$$p = P\left\{\frac{X_1 + \cdots + X_{25} - 1000}{20\sqrt{25}} > \frac{1100 - 1000}{20\sqrt{25}}\right\}$$
$$\approx P\{N(0, 1) > 1\}$$
$$= 1 - \Phi(1)$$
$$\approx 0.1587 \quad \blacksquare$$

We now present a heuristic proof of the Central Limit theorem. Suppose first that the X_i have mean 0 and variance 1, and let $E[e^{tX}]$ denote their common moment generating function. Then, the moment generating function of $\frac{X_1 + \cdots + X_n}{\sqrt{n}}$ is

$$E\left[\exp\left\{t\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)\right\}\right] = E\left[e^{tX_1/\sqrt{n}} e^{tX_2/\sqrt{n}} \cdots e^{tX_n/\sqrt{n}}\right]$$
$$= \left(E\left[e^{tX/\sqrt{n}}\right]\right)^n \qquad \text{by independence}$$

Now, for n large, we obtain from the Taylor series expansion of e^y that

$$e^{tX/\sqrt{n}} \approx 1 + \frac{tX}{\sqrt{n}} + \frac{t^2X^2}{2n}$$

Taking expectations shows that when n is large

$$E\left[e^{tX/\sqrt{n}}\right] \approx 1 + \frac{tE[X]}{\sqrt{n}} + \frac{t^2 E[X^2]}{2n}$$
$$= 1 + \frac{t^2}{2n} \qquad \text{because } E[X] = 0,\ E[X^2] = 1$$

Therefore, we obtain that when n is large

$$E\left[\exp\left\{t\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)\right\}\right] \approx \left(1 + \frac{t^2}{2n}\right)^n$$

When n goes to ∞ the approximation can be shown to become exact and we have that

$$\lim_{n\to\infty} E\left[\exp\left\{t\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)\right\}\right] = e^{t^2/2}$$

Thus, the moment generating function of $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converges to the moment generating function of a (standard) normal random variable with mean 0 and variance 1. Using this, it can be proven that the distribution function of the random variable $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converges to the standard normal distribution function Φ.

When the X_i have mean μ and variance σ^2, the random variables $\frac{X_i-\mu}{\sigma}$ have mean 0 and variance 1. Thus, the preceding shows that

$$P\left\{\frac{X_1 - \mu + X_2 - \mu + \cdots + X_n - \mu}{\sigma\sqrt{n}} \leqslant a\right\} \to \Phi(a)$$

which proves the central limit theorem.

2.8. Stochastic Processes

A *stochastic process* $\{X(t), t \in T\}$ is a collection of random variables. That is, for each $t \in T$, $X(t)$ is a random variable. The index t is often interpreted as time and, as a result, we refer to $X(t)$ as the *state* of the process at time t. For example, $X(t)$ might equal the total number of customers that have entered a supermarket by time t; or the number of customers in the supermarket at time t; or the total amount of sales that have been recorded in the market by time t; etc.

The set T is called the *index* set of the process. When T is a countable set the stochastic process is said to be a *discrete-time* process. If T is an interval of the real line, the stochastic process is said to be a *continuous-time* process. For instance, $\{X_n, n = 0, 1, \ldots\}$ is a discrete-time stochastic process indexed by the nonnegative integers; while $\{X(t), t \geqslant 0\}$ is a continuous-time stochastic process indexed by the nonnegative real numbers.

The *state space* of a stochastic process is defined as the set of all possible values that the random variables $X(t)$ can assume.

Thus, a stochastic process is a family of random variables that describes the evolution through time of some (physical) process. We shall see much of stochastic processes in the following chapters of this text.

Example 2.53 Consider a particle that moves along a set of $m + 1$ nodes, labeled $0, 1, \ldots, m$, that are arranged around a circle (see Figure 2.3). At each step the particle is equally likely to move one position in either the clockwise or counterclockwise direction. That is, if X_n is the position of the particle after its nth step then

$$P\{X_{n+1} = i + 1 | X_n = i\} = P\{X_{n+1} = i - 1 | X_n = i\} = \tfrac{1}{2}$$

where $i + 1 \equiv 0$ when $i = m$, and $i - 1 \equiv m$ when $i = 0$. Suppose now that the particle starts at 0 and continues to move around according to the preceding rules until all the nodes $1, 2, \ldots, m$ have been visited. What is the probability that node $i, i = 1, \ldots, m$, is the last one visited?

Solution: Surprisingly enough, the probability that node i is the last node visited can be determined without any computations. To do so, consider the first time that the particle is at one of the two neighbors of node i, that is, the first time that the particle is at one of the nodes $i - 1$ or $i + 1$ (with $m + 1 \equiv 0$). Suppose it is at node $i - 1$ (the argument in the alternative situation is identical). Since neither node i nor $i + 1$ has yet been visited, it follows that i will be the last node visited if and only if $i + 1$ is visited before i. This is so because in

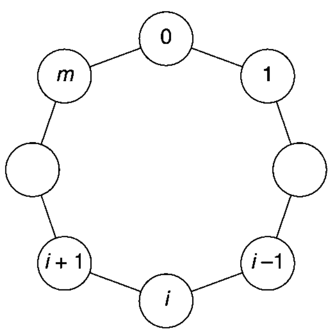

Figure 2.3. Particle moving around a circle.

order to visit $i+1$ before i the particle will have to visit all the nodes on the counterclockwise path from $i-1$ to $i+1$ before it visits i. But the probability that a particle at node $i-1$ will visit $i+1$ before i is just the probability that a particle will progress $m-1$ steps in a specified direction before progressing one step in the other direction. That is, it is equal to the probability that a gambler who starts with one unit, and wins one when a fair coin turns up heads and loses one when it turns up tails, will have his fortune go up by $m-1$ before he goes broke. Hence, because the preceding implies that the probability that node i is the last node visited is the same for all i, and because these probabilities must sum to 1, we obtain

$$P\{i \text{ is the last node visited}\} = 1/m, \qquad i = 1, \ldots, m \quad \blacksquare$$

Remark The argument used in Example 2.53 also shows that a gambler who is equally likely to either win or lose one unit on each gamble will be down n before being up 1 with probability $1/(n+1)$; or equivalently

$$P\{\text{gambler is up 1 before being down } n\} = \frac{n}{n+1}$$

Suppose now we want the probability that the gambler is up 2 before being down n. Upon conditioning on whether he reaches up 1 before down n, we obtain that

$$\begin{aligned}
&P\{\text{gambler is up 2 before being down } n\}\\
&\quad = P\{\text{up 2 before down } n | \text{up 1 before down } n\}\frac{n}{n+1}\\
&\quad = P\{\text{up 1 before down } n+1\}\frac{n}{n+1}\\
&\quad = \frac{n+1}{n+2}\frac{n}{n+1} = \frac{n}{n+2}
\end{aligned}$$

Repeating this argument yields that

$$P\{\text{gambler is up } k \text{ before being down } n\} = \frac{n}{n+k}$$

Exercises

1. An urn contains five red, three orange, and two blue balls. Two balls are randomly selected. What is the sample space of this experiment? Let X represent the number of orange balls selected. What are the possible values of X? Calculate $P\{X = 0\}$.

2. Let X represent the difference between the number of heads and the number of tails obtained when a coin is tossed n times. What are the possible values of X?

3. In Exercise 2, if the coin is assumed fair, then, for $n = 2$, what are the probabilities associated with the values that X can take on?

***4.** Suppose a die is rolled twice. What are the possible values that the following random variables can take on?

(i) The maximum value to appear in the two rolls.
(ii) The minimum value to appear in the two rolls.
(iii) The sum of the two rolls.
(iv) The value of the first roll minus the value of the second roll.

5. If the die in Exercise 4 is assumed fair, calculate the probabilities associated with the random variables in (i)–(iv).

6. Suppose five fair coins are tossed. Let E be the event that all coins land heads. Define the random variable I_E

$$I_E = \begin{cases} 1, & \text{if } E \text{ occurs} \\ 0, & \text{if } E^c \text{ occurs} \end{cases}$$

For what outcomes in the original sample space does I_E equal 1? What is $P\{I_E = 1\}$?

7. Suppose a coin having probability 0.7 of coming up heads is tossed three times. Let X denote the number of heads that appear in the three tosses. Determine the probability mass function of X.

8. Suppose the distribution function of X is given by

$$F(b) = \begin{cases} 0, & b < 0 \\ \frac{1}{2}, & 0 \leqslant b < 1 \\ 1, & 1 \leqslant b < \infty \end{cases}$$

What is the probability mass function of X?

9. If the distribution function of F is given by

$$F(b) = \begin{cases} 0, & b < 0 \\ \frac{1}{2}, & 0 \leqslant b < 1 \\ \frac{3}{5}, & 1 \leqslant b < 2 \\ \frac{4}{5}, & 2 \leqslant b < 3 \\ \frac{9}{10}, & 3 \leqslant b < 3.5 \\ 1, & b \geqslant 3.5 \end{cases}$$

calculate the probability mass function of X.