Probability_Ross_module_4_Notes

02 December 2022 14:25

Probability_Ross_module_4_Notes

CONTENTS

Chapter

4 RANDOM VARIABLES

Contents

4.1 Random Variables

When an experiment is performed, we are frequently interested mainly in some function of the outcome as opposed to the actual 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 separate values of each die. That is, we may be interested in knowing that the sum is 7 and may not be concerned over whether the actual outcome was (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), or (6, 1). Also, in flipping a coin, we may be interested in the total number of heads that occur and not care at all about the actual head–tail sequence that results. These quantities of interest, or, more formally, these real-valued functions defined on the sample space, are known as *random variables*.

Because 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 1a Suppose that our experiment consists of tossing 3 fair coins. If we let Y denote the number of heads that appear, then Y is a random variable taking on one of the values 0, 1, 2, and 3 with respective probabilities

$$P\{Y = 0\} = P\{(t,t,t)\} = \frac{1}{8}$$

$$P\{Y = 1\} = P\{(t,t,h),(t,h,t),(h,t,t)\} = \frac{3}{8}$$

$$P\{Y = 2\} = P\{(t,h,h),(h,t,h),(h,h,t)\} = \frac{3}{8}$$

$$P\{Y = 3\} = P\{(h,h,h)\} = \frac{1}{8}$$

Since Y must take on one of the values 0 through 3, we must have

$$1 = P\left(\bigcup_{i=0}^{3}\{Y = i\}\right) = \sum_{i=0}^{3} P\{Y = i\}$$

which, of course, is in accord with the preceding probabilities. ■

Example 1b

A life insurance agent has 2 elderly clients, each of whom has a life insurance policy that pays \$100,000 upon death. Let Y be the event that the younger one dies in the following year, and let O be the event that the older one dies in the following year. Assume that Y and O are independent, with respective probabilities $P(Y) = .05$ and $P(O) = .10$. If X denotes the total amount of money (in units of \$100,000) that will be paid out this year to any of these clients' beneficiaries, then X is a random variable that takes on one of the possible values 0, 1, 2 with respective probabilities

$$P\{X = 0\} = P(Y^c O^c) = P(Y^c)P(O^c) = (.95)(.9) = .855$$
$$P\{X = 1\} = P(YO^c) + P(Y^c O) = (.05)(.9) + (.95)(.1) = .140$$
$$P\{X = 2\} = P(YO) = (.05)(.1) = .005$$ ■

Example 1c

Four balls are to be randomly selected, without replacement, from an urn that contains 20 balls numbered 1 through 20. If X is the largest numbered ball selected, then X is a random variable that takes on one of the values $4, 5, \ldots, 20$. Because each of the $\binom{20}{4}$ possible selections of 4 of the 20 balls is equally likely, the probability that X takes on each of its possible values is

$$P\{X = i\} = \frac{\binom{i-1}{3}}{\binom{20}{4}}, \quad i = 4, \ldots, 20$$

This is so because the number of selections that result in $X = i$ is the number of selections that result in ball numbered i and three of the balls numbered 1 through $i - 1$ being selected. As there are $\binom{1}{1}\binom{i-1}{3}$ such selections, the preceding equation follows.

Suppose now that we want to determine $P\{X > 10\}$. One way, of course, is to just use the preceding to obtain

$$P\{X > 10\} = \sum_{i=11}^{20} P\{X = i\} = \sum_{i=11}^{20} \frac{\binom{i-1}{3}}{\binom{20}{4}}$$

However, a more direct approach for determining $P(X > 10)$ would be to use

$$P\{X > 10\} = 1 - P\{X \leq 10\} = 1 - \frac{\binom{10}{4}}{\binom{20}{4}}$$

where the preceding results because X will be less than or equal to 10 when the 4 balls chosen are among balls numbered 1 through 10. ■

Example 1d

Independent trials consisting of the flipping of a coin having probability p of coming up heads are continually performed until either a head occurs or a total of n flips is made. If we let X denote the number of times the coin is flipped, then X is a random variable taking on one of the values $1, 2, 3, \ldots, n$ with respective probabilities

$$P\{X = 1\} = P\{h\} = p$$

$$P\{X = 2\} = P\{(t,h)\} = (1 - p)p$$

$$P\{X = 3\} = P\{(t,t,h)\} = (1 - p)^2 p$$

$$\vdots$$

$$P\{X = n - 1\} = P\{(\underbrace{t,t,\ldots,t}_{n-2},h)\} = (1 - p)^{n-2}p$$

$$P\{X = n\} = P\{(\underbrace{t,t,\ldots,t}_{n-1},t),(\underbrace{t,t,\ldots,t}_{n-1},h)\} = (1 - p)^{n-1}$$

As a check, note that

$$\begin{aligned} P\left(\bigcup_{i=1}^{n}\{X = i\}\right) &= \sum_{i=1}^{n} P\{X = i\} \\ &= \sum_{i=1}^{n-1} p(1 - p)^{i-1} + (1 - p)^{n-1} \\ &= p\left[\frac{1 - (1 - p)^{n-1}}{1 - (1 - p)}\right] + (1 - p)^{n-1} \\ &= 1 - (1 - p)^{n-1} + (1 - p)^{n-1} \\ &= 1 \end{aligned}$$

■

Example 1e

Suppose that there are r distinct types of coupons and that each time one obtains a coupon, it is, independently of previous selections, equally likely to be any one of the r types. One random variable of interest is T, the number of coupons that need to be collected until one obtains a complete set of at least one of each type. Rather than derive $P\{T = n\}$ directly, let us start by considering the probability that T is greater than n. To do so, fix n and define the events $A_1, A_2, \ldots, A_r$ as follows: A_j is the event that no type j coupon is contained among the first n coupons collected, $j = 1, \ldots, r$. Hence, by the inclusion-exclusion identity

$$\begin{aligned} P\{T > n\} &= P\left(\bigcup_{j=1}^{r} A_j\right) \\ &= \sum_j P(A_j) - \sum_{j_1<j_2}\sum P(A_{j_1}A_{j_2}) + \cdots \\ &\quad + (-1)^{k+1}\sum_{j_1<j_2<\cdots<j_k}\sum\cdots\sum P(A_{j_1}A_{j_2}\cdots A_{j_k})\cdots \\ &\quad + (-1)^{r+1}P(A_1A_2\cdots A_r) \end{aligned}$$

Now, A_j will occur if each of the n coupons collected is not of type j. Since each of the coupons will not be of type j with probability $(r-1)/r$, we have, by the assumed independence of the types of successive coupons,

$$P(A_j) = \left(\frac{r-1}{r}\right)^n$$

Also, the event $A_{j_1}A_{j_2}$ will occur if none of the first n coupons collected is of either type j_1 or type j_2. Thus, again using independence, we see that

$$P(A_{j_1}A_{j_2}) = \left(\frac{r-2}{r}\right)^n$$

The same reasoning gives

$$P(A_{j_1}A_{j_2}\cdots A_{j_k}) = \left(\frac{r-k}{r}\right)^n$$

and we see that for $n > 0$,

$$\begin{aligned} P\{T > n\} &= r\left(\frac{r-1}{r}\right)^n - \binom{r}{2}\left(\frac{r-2}{r}\right)^n + \binom{r}{3}\left(\frac{r-3}{r}\right)^n - \cdots \\ &\quad + (-1)^r\binom{r}{r-1}\left(\frac{1}{r}\right)^n \\ &= \sum_{i=1}^{r-1}\binom{r}{i}\left(\frac{r-i}{r}\right)^n(-1)^{i+1} \end{aligned} \tag{1.1}$$

The probability that T equals n can now be obtained from the preceding formula by the use of

$$P\{T > n-1\} = P\{T = n\} + P\{T > n\}$$

or, equivalently,

$$P\{T = n\} = P\{T > n-1\} - P\{T > n\}$$

Another random variable of interest is the number of distinct types of coupons that are contained in the first n selections—call this random variable D_n. To compute $P\{D_n = k\}$, let us start by fixing attention on a particular set of k distinct types, and let us then determine the probability that this set constitutes the set of distinct types obtained in the first n selections. Now, in order for this to be the situation, it is necessary and sufficient that of the first n coupons obtained,

A: each is one of these k types

B: each of these k types is represented

Now, each coupon selected will be one of the k types with probability k/r, so the probability that A will be valid is $(k/r)^n$. Also, given that a coupon is of one of the k types under consideration, it is easy to see that it is equally likely to be of any one of these k types. Hence, the conditional probability of B given that A occurs is the same as the probability that a set of n coupons, each equally likely to be any of k possible types, contains a complete set of all k types. But this is just the probability that the number needed to amass a complete set, when choosing among k types, is less than or equal to n and is thus obtainable from Equation (1.1) with k replacing r. Thus, we have

$$P(A) = \left(\frac{k}{r}\right)^n$$

$$P(B|A) = 1 - \sum_{i=1}^{k-1} \binom{k}{i} \left(\frac{k-i}{k}\right)^n (-1)^{i+1}$$

Finally, as there are $\binom{r}{k}$ possible choices for the set of k types, we arrive at

$$\begin{aligned} P\{D_n = k\} &= \binom{r}{k} P(AB) \\ &= \binom{r}{k} \left(\frac{k}{r}\right)^n \left[1 - \sum_{i=1}^{k-1} \binom{k}{i} \left(\frac{k-i}{k}\right)^n (-1)^{i+1}\right] \end{aligned}$$

Remark We can obtain a useful bound on $P(T > n) = P(\cup_{j=1}^{r} A_j)$ by using Boole's inequality along with the inequality $e^{-x} \geq 1 - x$.

$$\begin{aligned} P(T > n) &= P(\cup_{j=1}^{r} A_j) \\ &\leq \sum_{j=1}^{r} P(A_j) \\ &= r(1 - \frac{1}{r})^n \\ &\leq re^{-n/r} \end{aligned}$$

The first inequality is Boole's inequality, which says that the probability of the union of events is always less than or equal to the sum of the probabilities of these events, and the last inequality uses that $e^{-1/r} \geq 1 - 1/r$. ■

For a random variable X, the function F defined by

$$F(x) = P\{X \leq x\} \qquad -\infty < x < \infty$$

is called the *cumulative distribution function* or, more simply, the *distribution function* of X. Thus, the distribution function specifies, for all real values x, the probability that the random variable is less than or equal to x.

Now, suppose that $a \leq b$. Then, because the event $\{X \leq a\}$ is contained in the event $\{X \leq b\}$, it follows that $F(a)$, the probability of the former, is less than or equal to $F(b)$, the probability of the latter. In other words, $F(x)$ is a nondecreasing function of x. Other general properties of the distribution function are given in Section 4.10.

4.2 Discrete Random Variables

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

$$p(x_i) \geq 0 \quad \text{for } i = 1, 2, \ldots$$
$$p(x) = 0 \quad \text{for all other values of } x$$

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

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

It is often instructive to present the probability mass function in a graphical format by plotting $p(x_i)$ on the y-axis against x_i on the x-axis. For instance, if the probability mass function of X is

$$p(0) = \frac{1}{4} \qquad p(1) = \frac{1}{2} \qquad p(2) = \frac{1}{4}$$

we can represent this function graphically as shown in Figure 4.1. Similarly, a graph of the probability mass function of the random variable representing the sum when two dice are rolled looks like Figure 4.2.

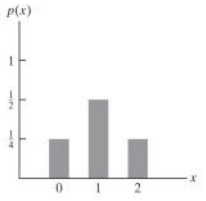


Figure 4.1

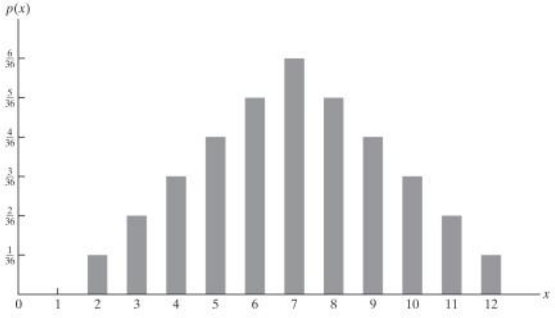


Figure 4.2

Example 2a

The probability mass function of a random variable X is given by $p(i) = c\lambda^i/i!$, $i = 0, 1, 2, \ldots$, where λ is some positive value. Find (a) $P\{X = 0\}$ and (b) $P\{X > 2\}$.

Solution Since $\sum_{i=0}^{\infty} p(i) = 1$, we have

$$c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = 1$$

which, because $e^x = \sum_{i=0}^{\infty} x^i/i!$, implies that

$$ce^{\lambda} = 1 \quad \text{or} \quad c = e^{-\lambda}$$

Hence,

(a) $P\{X = 0\} = e^{-\lambda}\lambda^0/0! = e^{-\lambda}$

(b) $P\{X > 2\} = 1 - P\{X \leq 2\} = 1 - P\{X = 0\} - P\{X = 1\} - P\{X = 2\}$

$$= 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2 e^{-\lambda}}{2}$$

■

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

$$F(a) = \sum_{\text{all } x \leq a} p(x)$$

If X is a discrete random variable whose possible values are $x_1, x_2, x_3, \ldots$, where $x_1 < x_2 < x_3 < \cdots$, then the distribution function F of X is a step function. That is, the value of F is constant in the intervals (x_{i-1}, x_i) and then takes a step (or jump) of size $p(x_i)$ at x_i. For instance, if X has a probability mass function given by

$$p(1) = \frac{1}{4} \qquad p(2) = \frac{1}{2} \qquad p(3) = \frac{1}{8} \qquad p(4) = \frac{1}{8}$$

then its cumulative distribution function is

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{4} & 1 \leq a < 2 \\ \frac{3}{4} & 2 \leq a < 3 \\ \frac{7}{8} & 3 \leq a < 4 \\ 1 & 4 \leq a \end{cases}$$

This function is depicted graphically in Figure 4.3.

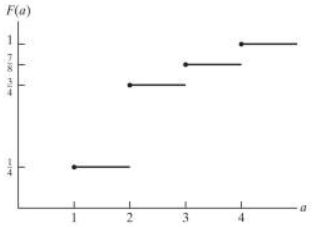


Figure 4.3

Note that the size of the step at any of the values 1, 2, 3, and 4 is equal to the probability that X assumes that particular value.

4.3 Expected Value

One of the most important concepts in probability theory is that of the expectation of a random variable. If X is a discrete random variable having a probability mass function $p(x)$, then the *expectation*, or the *expected value*, of X, denoted by $E[X]$, is defined by

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

In 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 it. For instance, on the one hand, if the probability mass function of X is given by

$$p(0) = \frac{1}{2} = p(1)$$

then

$$E[X] = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$$

is just the ordinary average of the two possible values, 0 and 1, that X can assume. On the other hand, if

$$p(0) = \frac{1}{3} \qquad p(1) = \frac{2}{3}$$

then

$$E[X] = 0\left(\frac{1}{3}\right) + 1\left(\frac{2}{3}\right) = \frac{2}{3}$$

is a weighted average of the two possible values 0 and 1, where the value 1 is given twice as much weight as the value 0, since $p(1) = 2p(0)$.

Another motivation of the definition of expectation is provided by the frequency interpretation of probabilities. This interpretation (partially justified by the strong law of large numbers, to be presented in Chapter 8) assumes that if an infinite sequence of independent replications of an experiment is performed, then, for any event E, the proportion of time that E occurs will be $P(E)$. Now, consider a random variable X that must take on one of the values $x_1, x_2, \ldots x_n$ with respective probabilities $p(x_1), p(x_2), \ldots, p(x_n)$, and think of X as representing our winnings in a single game of chance. That is, with probability $p(x_i)$, we shall win x_i units $i = 1, 2, \ldots, n$. By the frequency interpretation, if we play this game continually, then the proportion of time that we win x_i will be $p(x_i)$. Since this is true for all $i, i = 1, 2, \ldots, n$, it follows that our average winnings per game will be

$$\sum_{i=1}^{n} x_i p(x_i) = E[X]$$

Example 3a 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

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

Example 3b

We say that I is an indicator variable for the event A if

$$I = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A^c \text{ occurs} \end{cases}$$

Find $E[I]$.

Solution Since $p(1) = P(A)$, $p(0) = 1 - P(A)$, we have

$$E[I] = P(A)$$

That is, the expected value of the indicator variable for the event A is equal to the probability that A occurs. $\blacksquare$

Example 3c

A contestant on a quiz show is presented with two questions, questions 1 and 2, which he is to attempt to answer in some order he chooses. If he decides to try question i first, then he will be allowed to go on to question $j, j \neq i$, only if his answer to question i is correct. If his initial answer is incorrect, he is not allowed to answer the other question. The contestant is to receive V_i dollars if he answers question i correctly, $i = 1, 2$. For instance, he will receive $V_1 + V_2$ dollars if he answers both questions correctly. If the probability that he knows the answer to question i is $P_i, i = 1, 2$, which question should he attempt to answer first so as to maximize his expected winnings? Assume that the events $E_i, i = 1, 2$, that he knows the answer to question i are independent events.

Solution On the one hand, if he attempts to answer question 1 first, then he will win

$$\begin{array}{ll} 0 & \text{with probability } 1 - P_1 \\ V_1 & \text{with probability } P_1(1 - P_2) \\ V_1 + V_2 & \text{with probability } P_1P_2 \end{array}$$

Hence, his expected winnings in this case will be

$$V_1P_1(1 - P_2) + (V_1 + V_2)P_1P_2$$

On the other hand, if he attempts to answer question 2 first, his expected winnings will be

$$V_2P_2(1 - P_1) + (V_1 + V_2)P_1P_2$$

Therefore, it is better to try question 1 first if

$$V_1P_1(1 - P_2) \geq V_2P_2(1 - P_1)$$

or, equivalently, if

$$\frac{V_1P_1}{1 - P_1} \geq \frac{V_2P_2}{1 - P_2}$$

For example, if he is 60 percent certain of answering question 1, worth \$200, correctly and he is 80 percent certain of answering question 2, worth \$100, correctly, then he should attempt to answer question 2 first because

$$400 = \frac{(100)(.8)}{.2} > \frac{(200)(.6)}{.4} = 300$$ ■

Example 3d

A school class of 120 students is driven in 3 buses to a symphonic performance. There are 36 students in one of the buses, 40 in another, and 44 in the third bus. When the buses arrive, one of the 120 students is randomly chosen. Let X denote the number of students on the bus of that randomly chosen student, and find $E[X]$.

Solution Since the randomly chosen student is equally likely to be any of the 120 students, it follows that

$$P\{X = 36\} = \frac{36}{120} \quad P\{X = 40\} = \frac{40}{120} \quad P\{X = 44\} = \frac{44}{120}$$

Hence,

$$E[X] = 36\left(\frac{3}{10}\right) + 40\left(\frac{1}{3}\right) + 44\left(\frac{11}{30}\right) = \frac{1208}{30} = 40.2667$$

However, the average number of students on a bus is $120/3 = 40$, showing that the expected number of students on the bus of a randomly chosen student is larger than the average number of students on a bus. This is a general phenomenon, and it occurs because the more students there are on a bus, the more likely it is that a randomly chosen student would have been on that bus. As a result, buses with many students are given more weight than those with fewer students. (See Self-Test Problem 4.4) ■

Remark The probability concept of expectation is analogous to the physical concept of the *center of gravity* of a distribution of mass. Consider a discrete random variable X having probability mass function $p(x_i), i \geq 1$. If we now imagine a weightless rod in which weights with mass $p(x_i), i \geq 1$, are located at the points $x_i, i \geq 1$ (see Figure 4.4), then the point at which the rod would be in balance is known as the center of gravity. For those readers acquainted with elementary statics, it is now a simple matter to show that this point is at $E[X]$.[†] ■

$p(-1) = .10,\ p(0) = .25,\ p(1) = .30,\ p(2) = .35$

∧ = center of gravity = .9

Figure 4.4

4.4 Expectation of a Function of a Random Variable

Suppose that we are given a discrete random variable along with its probability mass function and that we want to compute the expected value of some function of X, say, $g(X)$. How can we accomplish this? One way is as follows: Since $g(X)$ is itself a discrete random variable, it has a probability mass function, which can be determined from the probability mass function of X. Once we have determined the probability

[†] To prove this, we must show that the sum of the torques tending to turn the point around $E[X]$ is equal to 0. That is, we must show that $0 = \sum_i (x_i - E[X])p(x_i)$, which is immediate.

mass function of $g(X)$, we can compute $E[g(X)]$ by using the definition of expected value.

Example 4a Let X denote a random variable that takes on any of the values -1, 0, and 1 with respective probabilities

$$P\{X = -1\} = .2 \quad P\{X = 0\} = .5 \quad P\{X = 1\} = .3$$

Compute $E[X^2]$.

Solution Let $Y = X^2$. Then the probability mass function of Y is given by

$$P\{Y = 1\} = P\{X = -1\} + P\{X = 1\} = .5$$
$$P\{Y = 0\} = P\{X = 0\} = .5$$

Hence,

$$E[X^2] = E[Y] = 1(.5) + 0(.5) = .5$$

Note that

$$.5 = E[X^2] \neq (E[X])^2 = .01$$ ■

Although the preceding procedure will always enable us to compute the expected value of any function of X from a knowledge of the probability mass function of X, there is another way of thinking about $E[g(X)]$: Since $g(X)$ will equal $g(x)$ whenever X is equal to x, it seems reasonable that $E[g(X)]$ should just be a weighted average of the values $g(x)$, with $g(x)$ being weighted by the probability that X is equal to x. That is, the following result is quite intuitive.

Proposition 4.1 If X is a discrete random variable that takes on one of the values x_i, $i \geq 1$, with respective probabilities $p(x_i)$, then, for any real-valued function g,

$$E[g(X)] = \sum_i g(x_i)p(x_i)$$

Before proving this proposition, let us check that it is in accord with the results of Example 4a. Applying it to that example yields

$$\begin{aligned} E[X^2] &= (-1)^2(.2) + 0^2(.5) + 1^2(.3) \\ &= 1(.2 + .3) + 0(.5) \\ &= .5 \end{aligned}$$

which is in agreement with the result given in Example 4a.

Proof of Proposition 4.1 The proof of Proposition 4.1 proceeds, as in the preceding verification, by grouping together all the terms in $\sum_i g(x_i)p(x_i)$ having the same value of $g(x_i)$. Specifically, suppose that y_j, $j \geq 1$, represent the different values of $g(x_i)$, $i \geq 1$. Then, grouping all the $g(x_i)$ having the same value gives

$$\begin{aligned}\sum_i g(x_i)p(x_i) &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p(x_i)\\ &= \sum_j \sum_{i:g(x_i)=y_j} y_j p(x_i)\\ &= \sum_j y_j \sum_{i:g(x_i)=y_j} p(x_i)\\ &= \sum_j y_j P\{g(X) = y_j\}\\ &= E[g(X)]\end{aligned}$$

□

Example 4b A product that is sold seasonally yields a net profit of b dollars for each unit sold and a net loss of ℓ dollars for each unit left unsold when the season ends. The number of units of the product that are ordered at a specific department store during any season is a random variable having probability mass function $p(i), i \geq 0$. If the store must stock this product in advance, determine the number of units the store should stock so as to maximize its expected profit.

Solution Let X denote the number of units ordered. If s units are stocked, then the profit—call it $P(s)$—can be expressed as

$$\begin{aligned}P(s) &= bX - (s - X)\ell \quad \text{if } X \leq s\\ &= sb \qquad\qquad\qquad \text{if } X > s\end{aligned}$$

Hence, the expected profit equals

$$\begin{aligned}E[P(s)] &= \sum_{i=0}^{s} [bi - (s - i)\ell]p(i) + \sum_{i=s+1}^{\infty} sbp(i)\\ &= (b + \ell)\sum_{i=0}^{s} ip(i) - s\ell\sum_{i=0}^{s} p(i) + sb\left[1 - \sum_{i=0}^{s} p(i)\right]\\ &= (b + \ell)\sum_{i=0}^{s} ip(i) - (b + \ell)s\sum_{i=0}^{s} p(i) + sb\\ &= sb + (b + \ell)\sum_{i=0}^{s} (i - s)p(i)\end{aligned}$$

To determine the optimum value of s, let us investigate what happens to the profit when we increase s by 1 unit. By substitution, we see that the expected profit in this case is given by

$$\begin{aligned}E[P(s + 1)] &= b(s + 1) + (b + \ell)\sum_{i=0}^{s+1} (i - s - 1)p(i)\\ &= b(s + 1) + (b + \ell)\sum_{i=0}^{s} (i - s - 1)p(i)\end{aligned}$$

Therefore,

$$E[P(s + 1)] - E[P(s)] = b - (b + \ell)\sum_{i=0}^{s} p(i)$$

Thus, stocking $s + 1$ units will be better than stocking s units whenever

$$\sum_{i=0}^{s} p(i) < \frac{b}{b + \ell} \tag{4.1}$$

Because the left-hand side of Equation (4.1) is increasing in s while the right-hand side is constant, the inequality will be satisfied for all values of $s \leq s^*$, where s^* is the largest value of s satisfying Equation (4.1). Since

$$E[P(0)] < \cdots < E[P(s^*)] < E[P(s^* + 1)] > E[P(s^* + 2)] > \cdots$$

it follows that stocking $s^* + 1$ items will lead to a maximum expected profit. ■

Example 4c

Utility

Suppose that you must choose one of two possible actions, each of which can result in any of n consequences, denoted as $C_1, \ldots, C_n$. Suppose that if the first action is chosen, then consequence C_i will result with probability $p_i, i = 1, \ldots, n$, whereas if the second action is chosen, then consequence C_i will result with probability q_i, $i = 1, \ldots, n$, where $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$. The following approach can be used to determine which action to choose: Start by assigning numerical values to the different consequences. First, identify the least and the most desirable consequences—call them c and C, respectively; give consequence c the value 0 and give C the value 1. Now consider any of the other $n - 2$ consequences, say, C_i. To value this consequence, imagine that you are given the choice between either receiving C_i or taking part in a random experiment that either earns you consequence C with probability u or consequence c with probability $1 - u$. Clearly, your choice will depend on the value of u. On the one hand, if $u = 1$, then the experiment is certain to result in consequence C, and since C is the most desirable consequence, you will prefer participating in the experiment to receiving C_i. On the other hand, if $u = 0$, then the experiment will result in the least desirable consequence—namely, c—so in this case you will prefer the consequence C_i to participating in the experiment. Now, as u decreases from 1 to 0, it seems reasonable that your choice will at some point switch from participating in the experiment to the certain return of C_i, and at that critical switch point you will be indifferent between the two alternatives. Take that indifference probability u as the value of the consequence C_i. In other words, the value of C_i is that probability u such that you are indifferent between either receiving the consequence C_i or taking part in an experiment that returns consequence C with probability u or consequence c with probability $1 - u$. We call this indifference probability the *utility* of the consequence C_i, and we designate it as $u(C_i)$.

To determine which action is superior, we need to evaluate each one. Consider the first action, which results in consequence C_i with probability $p_i, i = 1, \ldots, n$. We can think of the result of this action as being determined by a two-stage experiment. In the first stage, one of the values $1, \ldots, n$ is chosen according to the probabilities $p_1, \ldots, p_n$; if value i is chosen, you receive consequence C_i. However, since C_i is equivalent to obtaining consequence C with probability $u(C_i)$ or consequence c with probability $1 - u(C_i)$, it follows that the result of the two-stage experiment is equivalent to an experiment in which either consequence C or consequence c is obtained,

with C being obtained with probability

$$\sum_{i=1}^{n} p_i u(C_i)$$

Similarly, the result of choosing the second action is equivalent to taking part in an experiment in which either consequence C or consequence c is obtained, with C being obtained with probability

$$\sum_{i=1}^{n} q_i u(C_i)$$

Since C is preferable to c, it follows that the first action is preferable to the second action if

$$\sum_{i=1}^{n} p_i u(C_i) > \sum_{i=1}^{n} q_i u(C_i)$$

In other words, the worth of an action can be measured by the expected value of the utility of its consequence, and the action with the largest expected utility is the most preferable. ■

A simple logical consequence of Proposition 4.1 is Corollary 4.1.

Corollary 4.1 If a and b are constants, then

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

Proof

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

□

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 $E[X^n], n \geq 1$, is called the *nth moment of X*. By Proposition 4.1, we note that

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

4.5 Variance

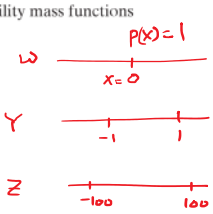
Given a random variable X along with its distribution function F, it would be extremely useful if we were able to summarize the essential properties of F by certain suitably defined measures. One such measure would be $E[X]$, the expected value of X. However, although $E[X]$ yields the weighted average of the possible values of X, it does not tell us anything about the variation, or spread, of these values. For

instance, although random variables W, Y, and Z having probability mass functions determined by

$$W = 0 \quad \text{with probability } 1$$

$$Y = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ +1 & \text{with probability } \frac{1}{2} \end{cases}$$

$$Z = \begin{cases} -100 & \text{with probability } \frac{1}{2} \\ +100 & \text{with probability } \frac{1}{2} \end{cases}$$



all have the same expectation—namely, 0—there is a much greater spread in the possible values of Y than in those of W (which is a constant) and in the possible values of Z than in those of Y.

Because we expect X to take on values around its mean $E[X]$, it would appear that a reasonable way of measuring the possible variation of X would be to look at how far apart X would be from its mean, on the average. One possible way to measure this variation would be to consider the quantity $E[|X - \mu|]$, where $\mu = E[X]$. However, it turns out to be mathematically inconvenient to deal with this quantity, so a more tractable quantity is usually considered—namely, the expectation of the square of the difference between X and its mean. We thus have the following definition.

Definition

If X is a random variable with mean μ, then the variance of X, denoted by $\text{Var}(X)$, is defined by

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

An alternative formula for $\text{Var}(X)$ is derived as follows:

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] \\ &= \sum_x (x - \mu)^2 p(x) \\ &= \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\ &= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned}$$

That is,

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

In words, the variance of X is equal to the expected value of X^2 minus the square of its expected value. In practice, this formula frequently offers the easiest way to compute $\text{Var}(X)$.

Example 5a Calculate $\text{Var}(X)$ if X represents the outcome when a fair die is rolled.

$E(W) = 0 \times 1 = 0$

$E(Y) = (-1)\frac{1}{2} + (1)\frac{1}{2} = 0$

$E(Z) = (-100)\frac{1}{2} + (100)\frac{1}{2} = 0$

Solution It was shown in Example 3a that $E[X] = \frac{7}{2}$. Also,

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

Hence,

$$\text{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$ ■

Because $\text{Var}(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 P(X = x)$ is the sum of nonnegative terms, it follows that $\text{Var}(X) \geq 0$ or equivalently, that

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

That is, the expected value of the square of a random variable is at least as large as the square of its expected value.

Example 5b

The *friendship paradox* is often expressed as saying that on average your friends have more friends than you do. More formally, suppose that there are n people in a certain population, labeled $1, 2, \ldots, n$, and that certain pairs of these individuals are friends. This *friendship network* can be graphically represented by having a circle for each person and then having a line between circles to indicate that those people are friends. For instance, Figure 4.5 indicates that there are 4 people in the community and that persons 1 and 2 are friends, persons 1 and 3 are friends, persons 1 and 4 are friends, and persons 2 and 4 are friends.

Let $f(i)$ denote the number of friends of person i and let $f = \sum_{i=1}^{n} f(i)$. (Thus, for the network of Figure 4.5, $f(1) = 3$, $f(2) = 2$, $f(3) = 1$, $f(4) = 2$ and $f = 8$.) Now, let X be a randomly chosen individual, equally likely to be any of $1, 2, \ldots, n$. That is,

$$P(X = i) = 1/n, \; i = 1, \ldots, n.$$

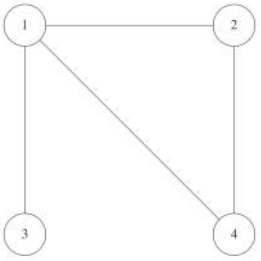


Figure 4.5 A Friendship Graph

Letting $g(i) = f(i)$ in Proposition 4.1, it follows that $E[f(X)]$, the expected number of friends of X, is

$$E[f(X)] = \sum_{i=1}^{n} f(i)P(X = i) = \sum_{i=1}^{n} f(i)/n = f/n$$

Also, letting $g(i) = f^2(i)$, it follows from Proposition 4.1 that $E[f^2(X)]$, the expected value of the square of the number of friends of X, is

$$E[f^2(X)] = \sum_{i=1}^{n} f^2(i)P(X = i) = \sum_{i=1}^{n} f^2(i)/n$$

Consequently, we see that

$$\frac{E[f^2(X)]}{E[f(X)]} = \frac{\sum_{i=1}^{n} f^2(i)}{f} \tag{5.1}$$

Now suppose that each of the n individuals writes the names of all their friends, with each name written on a separate sheet of paper. Thus, an individual with k friends will use k separate sheets. Because person i has $f(i)$ friends, there will be $f = \sum_{i=1}^{n} f(i)$ separate sheets of paper, with each sheet containing one of the n names. Now choose one of these sheets at random and let Y denote the name on that sheet. Let us compute $E[f(Y)]$, the expected number of friends of the person whose name is on the chosen sheet. Now, because person i has $f(i)$ friends, it follows that i is the name on $f(i)$ of the sheets, and thus i is the name on the chosen sheet with probability $\frac{f(i)}{f}$. That is,

$$P(Y = i) = \frac{f(i)}{f}, \quad i = 1, \ldots, n.$$

Consequently,

$$E[f(Y)] = \sum_{i=1}^{n} f(i)P(Y = i) = \sum_{i=1}^{n} f^2(i)/f \tag{5.2}$$

Thus, from (5.1), we see that

$$E[f(Y)] = \frac{E[f^2(X)]}{E[f(X)]} \geq E[f(X)]$$

where the inequality follows because the expected value of the square of any random variable is always at least as large as the square of its expectation. Thus, $E[f(X)] \leq E[f(Y)]$, which says that the average number of friends that a randomly chosen individual has is less than (or equal to if all the individuals have the same number of friends) the average number of friends of a randomly chosen friend.

Remark The intuitive reason for the friendship paradox is that X is equally likely to be any of the n individuals. On the other hand Y is chosen with a probability proportional to its number of friends; that is, the more friends an individual has the more likely that individual will be Y. Thus, Y is biased towards individuals with a large number of friends and so it is not surprising that the average number of friends that Y has is larger than the average number of friends that X has. ■

The following is a further example illustrating the usefulness of the inequality that the expected value of a square is at least as large as the square of the expected value.

Example 5c

Suppose there are m days in a year, and that each person is independently born on day r with probability $p_r, r = 1, \ldots, m, \sum_{r=1}^{m} p_r = 1$. Let $A_{i,j}$ be the event that persons i and j are born on the same day.

(a) Find $P(A_{1,3})$

(b) Find $P(A_{1,3}|A_{1,2})$

(c) Show $P(A_{1,3}|A_{1,2}) \geq P(A_{1,3})$

Solution

(a) Because the event that 1 and 3 have the same birthday is the union of the m mutually exclusive events that they were both born on day $r, r = 1, \ldots, m$, we have that

$$P(A_{1,3}) = \sum_r p_r^2.$$

(b) Using the definition of conditional probability we obtain that

$$P(A_{1,3}|A_{1,2}) = \frac{P(A_{1,2}A_{1,3})}{P(A_{1,2})} = \frac{\sum_r p_r^3}{\sum_r p_r^2}$$

where the preceding used that $A_{1,2}A_{1,3}$ is the union of the m mutually exclusive events that $1, 2, 3$ were all born on day $r, r = 1, \ldots, m$.

(c) It follows from parts (a) and (b) that $P(A_{1,3}|A_{1,2}) \geq P(A_{1,3})$ is equivalent to $\sum_r p_r^3 \geq (\sum_r p_r^2)^2$. To prove this inequality, let X be a random variable that is equal to p_r with probability p_r. That is, $P(X = p_r) = p_r, r = 1, \ldots, m$. Then

$$E[X] = \sum_r p_r P(X = p_r) = \sum_r p_r^2, \qquad E[X^2] = \sum_r p_r^2 P(X = p_r) = \sum_r p_r^3$$

and the result follows because $E[X^2] \geq (E[X])^2$.

Remark The intuitive reason for why part (c) is true is that if the "popular days" are the ones whose probabilities are relatively large, then knowing that 1 and 2 share the same birthday makes it more likely (than when we have no information) that the birthday of 1 is a popular day and that makes it more likely that 3 will have the same birthday as does 1.

■

A useful identity is that for any constants a and b,

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

To prove this equality, let $\mu = E[X]$ and note from Corollary 4.1 that $E[aX + b] = a\mu + b$. Therefore,

$$\begin{aligned}\text{Var}(aX + b) &= E[(aX + b - a\mu - b)^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2 E[(X - \mu)^2] \\ &= a^2 \text{Var}(X)\end{aligned}$$

Remarks (a) Analogous to the means being the center of gravity of a distribution of mass, the variance represents, in the terminology of mechanics, the moment of inertia.

(b) The square root of the $\text{Var}(X)$ is called the *standard deviation* of X, and we denote it by $\text{SD}(X)$. That is,

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

Discrete random variables are often classified according to their probability mass functions. In the next few sections, we consider some of the more common types.

4.6 The Bernoulli and Binomial Random Variables

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

$$\begin{aligned}p(0) &= P\{X = 0\} = 1 - p \\ p(1) &= P\{X = 1\} = p\end{aligned} \tag{6.1}$$

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

A random variable X is said to be a *Bernoulli random variable* (after the Swiss mathematician James Bernoulli) if its probability mass function is given by Equations (6.1) for some $p \in (0, 1)$.

Suppose now that n independent trials, each of which results in a success with probability p or 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). Thus, a Bernoulli random variable is just a binomial random variable with parameters $(1, p)$.

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

$$p(i) = \binom{n}{i} p^i (1 - p)^{n-i} \qquad i = 0, 1, \ldots, n \tag{6.2}$$

The validity of Equation (6.2) may be verified by first noting that the probability of any particular sequence of n outcomes containing i successes and $n - i$ failures is, by the assumed independence of trials, $p^i(1 - p)^{n-i}$. Equation (6.2) then follows, since there are $\binom{n}{i}$ different sequences of the n outcomes leading to i successes and $n - i$ failures. This perhaps can most easily be seen by noting that there are $\binom{n}{i}$

different choices of the i trials that result in successes. For instance, if $n = 4, i = 2$, then there are $\binom{4}{2} = 6$ ways in which the four trials can result in two successes, namely, any of the outcomes (s, s, f, f), (s, f, s, f), (s, f, f, s), (f, s, s, f), (f, s, f, s), and (f, f, s, s), where the outcome (s, s, f, f) means, for instance, that the first two trials are successes and the last two failures. Since each of these outcomes has probability $p^2(1 - p)^2$ of occurring, the desired probability of two successes in the four trials is $\binom{4}{2} p^2(1 - p)^2$.

Note that, by the binomial theorem, the probabilities sum to 1; that is,

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

Example 6a Five fair coins are flipped. If the outcomes are assumed independent, find the probability mass function of the number of heads obtained.

Solution If we let X equal the number of heads (successes) that appear, then X is a binomial random variable with parameters $\left(n = 5, p = \frac{1}{2}\right)$. Hence, by Equation (6.2),

$$P\{X = 0\} = \binom{5}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$P\{X = 1\} = \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4 = \frac{5}{32}$$

$$P\{X = 2\} = \binom{5}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = \frac{10}{32}$$

$$P\{X = 3\} = \binom{5}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = \frac{10}{32}$$

$$P\{X = 4\} = \binom{5}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 = \frac{5}{32}$$

$$P\{X = 5\} = \binom{5}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 = \frac{1}{32}$$

■

Example 6b It is known that screws produced by a certain company will be defective with probability .01, independently of one another. The company sells the screws in packages of 10 and offers a money-back guarantee that at most 1 of the 10 screws is defective. What proportion of packages sold must the company replace?

Solution If X is the number of defective screws in a package, then X is a binomial random variable with parameters (10, .01). Hence, the probability that a package will have to be replaced is

$$1 - P\{X = 0\} - P\{X = 1\} = 1 - \binom{10}{0} (.01)^0 (.99)^{10} - \binom{10}{1} (.01)^1 (.99)^9$$
$$\approx .004$$

Thus, only .4 percent of the packages will have to be replaced. ■

Example 6c

The following gambling game, known as the wheel of fortune (or chuck-a-luck), is quite popular at many carnivals and gambling casinos: A player bets on one of the numbers 1 through 6. Three dice are then rolled, and if the number bet by the player appears i times, $i = 1, 2, 3$, then the player wins i units; if the number bet by the player does not appear on any of the dice, then the player loses 1 unit. Is this game fair to the player? (Actually, the game is played by spinning a wheel that comes to rest on a slot labeled by three of the numbers 1 through 6, but this variant is mathematically equivalent to the dice version.)

Solution If we assume that the dice are fair and act independently of one another, then the number of times that the number bet appears is a binomial random variable with parameters $\left(3, \frac{1}{6}\right)$. Hence, letting X denote the player's winnings in the game, we have

$$P\{X = -1\} = \binom{3}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3 = \frac{125}{216}$$

$$P\{X = 1\} = \binom{3}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^2 = \frac{75}{216}$$

$$P\{X = 2\} = \binom{3}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^1 = \frac{15}{216}$$

$$P\{X = 3\} = \binom{3}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0 = \frac{1}{216}$$

In order to determine whether or not this is a fair game for the player, let us calculate $E[X]$. From the preceding probabilities, we obtain

$$E[X] = \frac{-125 + 75 + 30 + 3}{216} = \frac{-17}{216}$$

Hence, in the long run, the player will lose 17 units per every 216 games he plays. ■

In the next example, we consider the simplest form of the theory of inheritance as developed by Gregor Mendel (1822–1884).

Example 6d

Suppose that a particular trait (such as eye color or left-handedness) of a person is classified on the basis of one pair of genes, and suppose also that d represents a dominant gene and r a recessive gene. Thus, a person with dd genes is purely dominant, one with rr is purely recessive, and one with rd is hybrid. The purely dominant and the hybrid individuals are alike in appearance. Children receive 1 gene from each parent. If, with respect to a particular trait, 2 hybrid parents have a total of 4 children, what is the probability that 3 of the 4 children have the outward appearance of the dominant gene?

The preceding Figure 4.6a and b shows what can happen when hybrid yellow (dominant) and green (recessive) seeds are crossed.

Solution If we assume that each child is equally likely to inherit either of 2 genes from each parent, the probabilities that the child of 2 hybrid parents will have dd, rr, and rd pairs of genes are, respectively, $\frac{1}{4}, \frac{1}{4}$, and $\frac{1}{2}$. Hence, since an offspring will have the outward appearance of the dominant gene if its gene pair is either dd or rd,

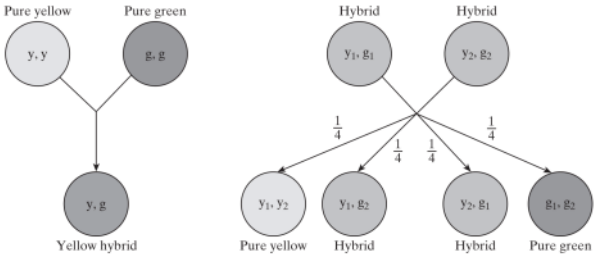


Figure 4.6 (a) Crossing pure yellow seeds with pure green seeds; (b) Crossing hybrid first-generation seeds.

it follows that the number of such children is binomially distributed with parameters $\left(4, \frac{3}{4}\right)$. Thus, the desired probability is

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

Example 6e

Consider a jury trial in which it takes 8 of the 12 jurors to convict the defendant; that is, in order for the defendant to be convicted, at least 8 of the jurors must vote him guilty. If we assume that jurors act independently and that whether or not the defendant is guilty, each makes the right decision with probability θ, what is the probability that the jury renders a correct decision?

Solution The problem, as stated, is incapable of solution, for there is not yet enough information. For instance, if the defendant is innocent, the probability of the jury rendering a correct decision is

$$\sum_{i=5}^{12}\binom{12}{i}\theta^i(1-\theta)^{12-i}$$

whereas, if he is guilty, the probability of a correct decision is

$$\sum_{i=8}^{12}\binom{12}{i}\theta^i(1-\theta)^{12-i}$$

Therefore, if α represents the probability that the defendant is guilty, then, by conditioning on whether or not he is guilty, we obtain the probability that the jury renders a correct decision:

$$\alpha\sum_{i=8}^{12}\binom{12}{i}\theta^i(1-\theta)^{12-i} + (1-\alpha)\sum_{i=5}^{12}\binom{12}{i}\theta^i(1-\theta)^{12-i}$$ ■

Example 6f

A communication system consists of n components, each of which will, independently, function with probability p. The total system will be able to operate effectively if at least one-half of its components function.

(a) For what values of p is a 5-component system more likely to operate effectively than a 3-component system?

(b) In general, when is a $(2k + 1)$-component system better than a $(2k - 1)$-component system?

Solution (a) Because the number of functioning components is a binomial random variable with parameters (n, p), it follows that the probability that a 5-component system will be effective is

$$\binom{5}{3}p^3(1 - p)^2 + \binom{5}{4}p^4(1 - p) + p^5$$

whereas the corresponding probability for a 3-component system is

$$\binom{3}{2}p^2(1 - p) + p^3$$

Hence, the 5-component system is better if

$$10p^3(1 - p)^2 + 5p^4(1 - p) + p^5 > 3p^2(1 - p) + p^3$$

which reduces to

$$3(p - 1)^2(2p - 1) > 0$$

or

$$p > \frac{1}{2}$$

(b) In general, a system with $2k + 1$ components will be better than one with $2k - 1$ components if (and only if) $p > \frac{1}{2}$. To prove this, consider a system of $2k + 1$ components and let X denote the number of the first $2k - 1$ that function. Then

$$\begin{aligned} P_{2k+1}(\text{effective}) = {} & P\{X \geq k + 1\} + P\{X = k\}(1 - (1 - p)^2) \\ & + P\{X = k - 1\}p^2 \end{aligned}$$

which follows because the $(2k + 1)$-component system will be effective if either

(i) $X \geq k + 1$;
(ii) $X = k$ and at least one of the remaining 2 components function; or
(iii) $X = k - 1$ and both of the next 2 components function.

Since

$$\begin{aligned} P_{2k-1}(\text{effective}) &= P\{X \geq k\} \\ &= P\{X = k\} + P\{X \geq k + 1\} \end{aligned}$$

we obtain

$$\begin{aligned}
&P_{2k+1}(\text{effective}) - P_{2k-1}(\text{effective}) \\
&\quad = P\{X = k-1\}p^2 - (1-p)^2 P\{X = k\} \\
&\quad = \binom{2k-1}{k-1} p^{k-1}(1-p)^k p^2 - (1-p)^2 \binom{2k-1}{k} p^k (1-p)^{k-1} \\
&\quad = \binom{2k-1}{k} p^k (1-p)^k [p - (1-p)] \text{ since } \binom{2k-1}{k-1} = \binom{2k-1}{k} \\
&\quad > 0 \Leftrightarrow p > \frac{1}{2}
\end{aligned}$$

■

4.6.1 Properties of Binomial Random Variables

We will now examine the properties of a binomial random variable with parameters n and p. To begin, let us compute its expected value and variance. To begin, note that

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

Using the identity

$$i\binom{n}{i} = n\binom{n-1}{i-1}$$

gives

$$\begin{aligned}
E[X^k] &= np \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1}(1-p)^{n-i} \\
&= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \quad \begin{matrix}\text{by letting} \\ j = i-1\end{matrix} \\
&= npE[(Y+1)^{k-1}]
\end{aligned}$$

where Y is a binomial random variable with parameters $n-1, p$. Setting $k = 1$ in the preceding equation yields

$$E[X] = np$$

That is, the expected number of successes that occur in n independent trials when each is a success with probability p is equal to np. Setting $k = 2$ in the preceding equation and using the preceding formula for the expected value of a binomial random variable yields

$$\begin{aligned}
E[X^2] &= npE[Y+1] \\
&= np[(n-1)p + 1]
\end{aligned}$$

Since $E[X] = np$, we obtain

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

Summing up, we have shown the following:

If X is a binomial random variable with parameters n and p, then

$$\begin{aligned}E[X] &= np \\ \text{Var}(X) &= np(1-p)\end{aligned}$$

The following proposition details how the binomial probability mass function first increases and then decreases.

Proposition 6.1 If X is a binomial random variable with parameters (n, p), where $0 < p < 1$, then as k goes from 0 to n, $P\{X = k\}$ first increases monotonically and then decreases monotonically, reaching its largest value when k is the largest integer less than or equal to $(n+1)p$.

Proof We prove the proposition by considering $P\{X = k\}/P\{X = k-1\}$ and determining for what values of k it is greater or less than 1. Now,

$$\frac{P\{X=k\}}{P\{X=k-1\}} = \frac{\frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!}p^{k-1}(1-p)^{n-k+1}}$$

$$= \frac{(n-k+1)p}{k(1-p)}$$

Hence, $P\{X = k\} \geq P\{X = k-1\}$ if and only if

$$(n-k+1)p \geq k(1-p)$$

or, equivalently, if and only if

$$k \leq (n+1)p$$

and the proposition is proved. □

As an illustration of Proposition 6.1, consider Figure 4.7, the graph of the probability mass function of a binomial random variable with parameters $(10, \frac{1}{2})$.

Example 6g In a U.S. presidential election, the candidate who gains the maximum number of votes in a state is awarded the total number of electoral college votes allocated to that state. The number of electoral college votes of a given state is roughly proportional to the population of that state—that is, a state with population n has roughly nc electoral votes. (Actually, it is closer to $nc + 2$, as a state is given an electoral vote for each member it has in the House of Representatives, with the number of such representatives being roughly proportional to the population of the state, and one electoral college vote for each of its two senators.) Let us determine the average power of a citizen in a state of size n in a close presidential election, where, by *average power in a close election*, we mean that a voter in a state of size $n = 2k + 1$ will be decisive if the other $n - 1$ voters split their votes evenly between the two candidates. (We are assuming here that n is odd, but the case where n is even is quite similar.)

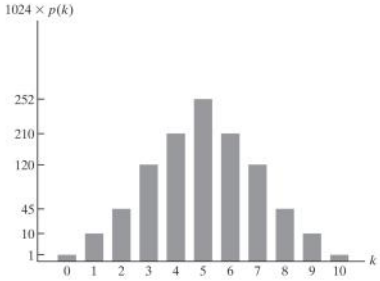


Figure 4.7 Graph of $p(k) = \binom{10}{k}\left(\frac{1}{2}\right)^{10}$.

Because the election is close, we shall suppose that each of the other $n - 1 = 2k$ voters acts independently and is equally likely to vote for either candidate. Hence, the probability that a voter in a state of size $n = 2k + 1$ will make a difference to the outcome is the same as the probability that $2k$ tosses of a fair coin land heads and tails an equal number of times. That is,

$$\begin{aligned} &P\{\text{voter in state of size } 2k + 1 \text{ makes a difference}\} \\ &\quad = \binom{2k}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^k \\ &\quad = \frac{(2k)!}{k!k!2^{2k}} \end{aligned}$$

To approximate the preceding equality, we make use of Stirling's approximation, which says that for k large,

$$k! \sim k^{k+1/2}e^{-k}\sqrt{2\pi}$$

where we say that $a_k \sim b_k$ when the ratio a_k/b_k approaches 1 as k approaches ∞. Hence, it follows that

$$\begin{aligned} &P\{\text{voter in state of size } 2k + 1 \text{ makes a difference}\} \\ &\quad \sim \frac{(2k)^{2k+1/2}e^{-2k}\sqrt{2\pi}}{k^{2k+1}e^{-2k}(2\pi)2^{2k}} = \frac{1}{\sqrt{k\pi}} \end{aligned}$$

Because such a voter (if he or she makes a difference) will affect nc electoral votes, the expected number of electoral votes a voter in a state of size n will affect—or the voter's average power—is given by

$$\begin{aligned} \text{average power} &= ncP\{\text{makes a difference}\} \\ &\sim \frac{nc}{\sqrt{n\pi/2}} \\ &= c\sqrt{2n/\pi} \end{aligned}$$

Thus, the average power of a voter in a state of size n is proportional to the square root of n, showing that in presidential elections, voters in large states have more power than do those in smaller states. ■

4.6.2 Computing the Binomial Distribution Function

Suppose that X is binomial with parameters (n, p). The key to computing its distribution function

$$P\{X \leq i\} = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k} \qquad i = 0, 1, \ldots, n$$

is to utilize the following relationship between $P\{X = k+1\}$ and $P\{X = k\}$, which was established in the proof of Proposition 6.1:

$$P\{X = k+1\} = \frac{p}{1-p} \frac{n-k}{k+1} P\{X = k\} \tag{6.3}$$

Example 6h

Let X be a binomial random variable with parameters $n = 6$, $p = .4$. Then, starting with $P\{X = 0\} = (.6)^6$ and recursively employing Equation (6.3), we obtain

$$P\{X = 0\} = (.6)^6 \approx .0467$$

$$P\{X = 1\} = \frac{4}{6}\frac{6}{1} P\{X = 0\} \approx .1866$$

$$P\{X = 2\} = \frac{4}{6}\frac{5}{2} P\{X = 1\} \approx .3110$$

$$P\{X = 3\} = \frac{4}{6}\frac{4}{3} P\{X = 2\} \approx .2765$$

$$P\{X = 4\} = \frac{4}{6}\frac{3}{4} P\{X = 3\} \approx .1382$$

$$P\{X = 5\} = \frac{4}{6}\frac{2}{5} P\{X = 4\} \approx .0369$$

$$P\{X = 6\} = \frac{4}{6}\frac{1}{6} P\{X = 5\} \approx .0041$$ ■

A computer program that utilizes the recursion (6.3) to compute the binomial distribution function is easily written. To compute $P\{X \leq i\}$, the program should first compute $P\{X = i\}$ and then use the recursion to successively compute $P\{X = i-1\}, P\{X = i-2\}$, and so on.

Historical note

Independent trials having a common probability of success p were first studied by the Swiss mathematician Jacques Bernoulli (1654–1705). In his book *Ars Conjectandi (The Art of Conjecturing)*, published by his nephew Nicholas eight years after his death in 1713, Bernoulli showed that if the number of such trials were large, then the proportion of them that were successes would be close to p with a probability near 1.

Jacques Bernoulli was from the first generation of the most famous mathematical family of all time. Altogether, there were between 8 and 12 Bernoullis, spread over three generations, who made fundamental contributions to probability, statistics, and mathematics. One difficulty in knowing their exact number is the fact that several had the same name. (For example, two of the sons of Jacques's brother Jean were named Jacques and Jean.) Another difficulty is that several of the Bernoullis were known by different names in different places. Our Jacques (sometimes written Jaques) was, for instance, also known as Jakob (sometimes written Jacob) and as James Bernoulli. But whatever their number, their influence and output were prodigious. Like the Bachs of music, the Bernoullis of mathematics were a family for the ages!

Example 6i If X is a binomial random variable with parameters $n = 100$ and $p = .75$, find $P\{X = 70\}$ and $P\{X \leq 70\}$.

Solution A binomial calculator can be used to obtain the following solutions:

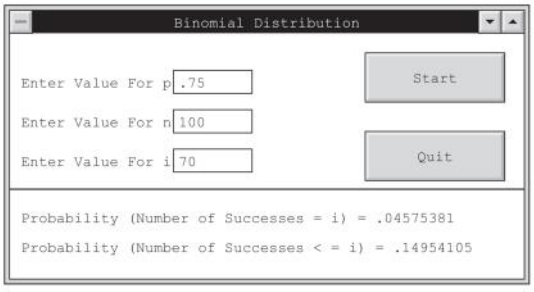


Figure 4.8

4.7 The Poisson Random Variable

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

$$p(i) = P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!} \quad i = 0, 1, 2, \ldots \tag{7.1}$$

Equation (7.1) 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 probability distribution was introduced by Siméon Denis Poisson in a book he wrote regarding the application of probability theory to lawsuits, criminal trials, and the like. This book, published in 1837, was entitled *Recherches sur la probabilité des jugements en matière criminelle et en matière civile (Investigations into the Probability of Verdicts in Criminal and Civil Matters)*.

The Poisson random variable has a tremendous range of applications in diverse areas because it may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small enough so that np is of moderate size. 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 λ moderate,

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

Hence, for n large and λ moderate,

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

In other words, if n independent trials, each of which results in a success with probability p, are performed, then when n is large and p is small enough to make np moderate, the number of successes occurring is approximately a Poisson random variable with parameter $\lambda = np$. This value λ (which will later be shown to equal the expected number of successes) will usually be determined empirically.

Some examples of random variables that generally obey the Poisson probability law [that is, they obey Equation (7.1)] are as follows:

1. The number of misprints on a page (or a group of pages) of a book
2. The number of people in a community who survive to age 100
3. The number of wrong telephone numbers that are dialed in a day
4. The number of packages of dog biscuits sold in a particular store each day
5. The number of customers entering a post office on a given day
6. The number of vacancies occurring during a year in the federal judicial system
7. The number of α-particles discharged in a fixed period of time from some radioactive material

Each of the preceding and numerous other random variables are approximately Poisson for the same reason—namely, because of the Poisson approximation to the binomial. For instance, we can suppose that there is a small probability p that each letter typed on a page will be misprinted. Hence, the number of misprints on a page will be approximately Poisson with $\lambda = np$, where n is the number of letters on a page. Similarly, we can suppose that each person in a community has some small probability of reaching age 100. Also, each person entering a store may be thought of as having some small probability of buying a package of dog biscuits, and so on.

Example 7a Suppose that the number of typographical errors on a single page of this book has a Poisson distribution with parameter $\lambda = \frac{1}{2}$. Calculate the probability that there is at least one error on this page.

Solution Letting X denote the number of errors on this page, we have

$$P\{X \geq 1\} = 1 - P\{X = 0\} = 1 - e^{-1/2} \approx .393 \quad \blacksquare$$

Example 7b Suppose that the probability that an item produced by a certain machine will be defective is .1. Find the probability that a sample of 10 items will contain at most 1 defective item.

Solution The desired probability is $\binom{10}{0}(.1)^0(.9)^{10} + \binom{10}{1}(.1)^1(.9)^9 = .7361$, whereas the Poisson approximation yields the value $e^{-1} + e^{-1} \approx .7358$. $\blacksquare$

Example 7c Consider an experiment that consists of counting the number of α particles given off in a 1-second interval by 1 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 2 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 of $3.2/n$ 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

$$\begin{aligned} P\{X \leq 2\} &= e^{-3.2} + 3.2e^{-3.2} + \frac{(3.2)^2}{2}e^{-3.2} \\ &\approx .3799 \end{aligned} \quad \blacksquare$$

Before computing the expected value and variance of the Poisson random variable with parameter λ, recall that this random variable approximates a binomial random variable with parameters n and p when n is large, p is small, and $\lambda = np$. Since such a binomial random variable has expected value $np = \lambda$ and variance $np(1 - p) = \lambda(1 - p) \approx \lambda$ (since p is small), it would seem that both the expected value and the variance of a Poisson random variable would equal its parameter λ. We now verify this result:

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} \\ &= \lambda \sum_{i=1}^{\infty} \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!} \\ &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \qquad \text{by letting } j = i - 1 \\ &= \lambda \qquad \text{since } \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda} \end{aligned}$$

Thus, the expected value of a Poisson random variable X is indeed equal to its parameter λ. To determine its variance, we first compute $E[X^2]$:

$$\begin{aligned} E[X^2] &= \sum_{i=0}^{\infty} \frac{i^2 e^{-\lambda}\lambda^i}{i!} \\ &= \lambda \sum_{i=1}^{\infty} \frac{i e^{-\lambda}\lambda^{i-1}}{(i-1)!} \\ &= \lambda \sum_{j=0}^{\infty} \frac{(j+1)e^{-\lambda}\lambda^j}{j!} \quad \text{by letting } j = i-1 \\ &= \lambda \left[\sum_{j=0}^{\infty} \frac{j e^{-\lambda}\lambda^j}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \right] \\ &= \lambda(\lambda + 1) \end{aligned}$$

where the final equality follows because the first sum is the expected value of a Poisson random variable with parameter λ and the second is the sum of the probabilities of this random variable. Therefore, since we have shown that $E[X] = \lambda$, we obtain

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= \lambda \end{aligned}$$

Hence, the expected value and variance of a Poisson random variable are both equal to its parameter λ.

We have shown that the Poisson distribution with parameter np is a very good approximation to the distribution of the number of successes in n independent trials when each trial has probability p of being a success, provided that n is large and p small. In fact, it remains a good approximation even when the trials are not independent, provided that their dependence is weak. For instance, recall the matching problem (Example 5m of Chapter 2) in which n men randomly select hats from a set consisting of one hat from each person. From the point of view of the number of men who select their own hat, we may regard the random selection as the result of n trials where we say that trial i is a success if person i selects his own hat, $i = 1, \ldots, n$. Defining the events $E_i, i = 1, \ldots, n$, by

$$E_i = \{\text{trial } i \text{ is a success}\}$$

it is easy to see that

$$P\{E_i\} = \frac{1}{n} \quad \text{and} \quad P\{E_i|E_j\} = \frac{1}{n-1}, \quad j \neq i$$

Thus, we see that although the events $E_i, i = 1, \ldots, n$ are not independent, their dependence, for large n, appears to be weak. Because of this, it seems reasonable to expect that the number of successes will approximately have a Poisson distribution with parameter $n \times 1/n = 1$ and indeed this is verified in Example 5m of Chapter 2.

For a second illustration of the strength of the Poisson approximation when the trials are weakly dependent, let us consider again the birthday problem presented in Example 5i of Chapter 2. In this example, we suppose that each of n people is equally likely to have any of the 365 days of the year as his or her birthday, and the problem

is to determine the probability that a set of n independent people all have different birthdays. A combinatorial argument was used to determine this probability, which was shown to be less than $\frac{1}{2}$ when $n = 23$.

We can approximate the preceding probability by using the Poisson approximation as follows: Imagine that we have a trial for each of the $\binom{n}{2}$ pairs of individuals i and $j, i \neq j$, and say that trial i, j is a success if persons i and j have the same birthday. If we let E_{ij} denote the event that trial i, j is a success, then, whereas the $\binom{n}{2}$ events $E_{ij}, 1 \leq i < j \leq n$, are not independent (see Theoretical Exercise 4.21), their dependence appears to be rather weak. (Indeed, these events are even *pairwise independent*, in that any 2 of the events E_{ij} and E_{kl} are independent—again, see Theoretical Exercise 4.21). Since $P(E_{ij}) = 1/365$, it is reasonable to suppose that the number of successes should approximately have a Poisson distribution with mean $\binom{n}{2} \Big/ 365 = n(n-1)/730$. Therefore,

$$\begin{aligned} P\{\text{no 2 people have the same birthday}\} &= P\{0 \text{ successes}\} \\ &\approx \exp\left\{\frac{-n(n-1)}{730}\right\} \end{aligned}$$

To determine the smallest integer n for which this probability is less than $\frac{1}{2}$, note that

$$\exp\left\{\frac{-n(n-1)}{730}\right\} \leq \frac{1}{2}$$

is equivalent to

$$\exp\left\{\frac{n(n-1)}{730}\right\} \geq 2$$

Taking logarithms of both sides, we obtain

$$\begin{aligned} n(n-1) &\geq 730 \log 2 \\ &\approx 505.997 \end{aligned}$$

which yields the solution $n = 23$, in agreement with the result of Example 5i of Chapter 2.

Suppose now that we wanted the probability that among the n people, no 3 of them have their birthday on the same day. Whereas this now becomes a difficult combinatorial problem, it is a simple matter to obtain a good approximation. To begin, imagine that we have a trial for each of the $\binom{n}{3}$ triplets i, j, k, where $1 \leq i < j < k \leq n$, and call the i, j, k trial a success if persons i, j, and k all have their birthday on the same day. As before, we can then conclude that the number of successes is approximately a Poisson random variable with parameter

$$\begin{aligned} \binom{n}{3} P\{i, j, k \text{ have the same birthday}\} &= \binom{n}{3}\left(\frac{1}{365}\right)^2 \\ &= \frac{n(n-1)(n-2)}{6 \times (365)^2} \end{aligned}$$

Hence,

$$P\{\text{no 3 have the same birthday}\} \approx \exp\left\{\frac{-n(n-1)(n-2)}{799350}\right\}$$

This probability will be less than $\frac{1}{2}$ when n is such that

$$n(n-1)(n-2) \geq 799350 \log 2 \approx 554067.1$$

which is equivalent to $n \geq 84$. Thus, the approximate probability that at least 3 people in a group of size 84 or larger will have the same birthday exceeds $\frac{1}{2}$.

For the number of events to occur to approximately have a Poisson distribution, it is not essential that all the events have the same probability of occurrence, but only that all of these probabilities be small. The following is referred to as the *Poisson paradigm*.

Poisson Paradigm. Consider n events, with p_i equal to the probability that event i occurs, $i = 1, \ldots, n$. If all the p_i are "small" and the trials are either independent or at most "weakly dependent," then the number of these events that occur approximately has a Poisson distribution with mean $\sum_{i=1}^{n} p_i$.

Our next example not only makes use of the Poisson paradigm, but also illustrates a variety of the techniques we have studied so far.

Example 7d **Length of the longest run**

A coin is flipped n times. Assuming that the flips are independent, with each one coming up heads with probability p, what is the probability that there is a string of k consecutive heads?

Solution We will first use the Poisson paradigm to approximate this probability. Now, if for $i = 1, \ldots, n-k+1$, we let H_i denote the event that flips $i, i+1, \ldots, i+k-1$ all land on heads, then the desired probability is that at least one of the events H_i occur. Because H_i is the event that starting with flip i, the next k flips all land on heads, it follows that $P(H_i) = p^k$. Thus, when p^k is small, we might think that the number of the H_i that occur should have an approximate Poisson distribution. However, such is not the case, because, although the events all have small probabilities, some of their dependencies are too great for the Poisson distribution to be a good approximation. For instance, because the conditional probability that flips $2, \ldots, k+1$ are all heads given that flips $1, \ldots, k$ are all heads is equal to the probability that flip $k+1$ is a head, it follows that

$$P(H_2|H_1) = p$$

which is far greater than the unconditional probability of H_2.

The trick that enables us to use a Poisson approximation is to note that there will be a string of k consecutive heads either if there is such a string that is immediately followed by a tail or if the final k flips all land on heads. Consequently, for $i = 1, \ldots, n-k$, let E_i be the event that flips $i, \ldots, i+k-1$ are all heads and flip $i+k$ is a tail; also, let E_{n-k+1} be the event that flips $n-k+1, \ldots, n$ are all heads. Note that

$$P(E_i) = p^k(1 - p), \quad i \leq n - k$$
$$P(E_{n-k+1}) = p^k$$

Thus, when p^k is small, each of the events E_i has a small probability of occurring. Moreover, for $i \neq j$, if the events E_i and E_j refer to nonoverlapping sequences of flips, then $P(E_i|E_j) = P(E_i)$; if they refer to overlapping sequences, then $P(E_i|E_j) = 0$. Hence, in both cases, the conditional probabilities are close to the unconditional ones, indicating that N, the number of the events E_i that occur, should have an approximate Poisson distribution with mean

$$E[N] = \sum_{i=1}^{n-k+1} P(E_i) = (n - k)p^k(1 - p) + p^k$$

Because there will not be a run of k heads if (and only if) $N = 0$, the preceding gives

$$P(\text{no head strings of length } k) = P(N = 0) \approx \exp\{-(n - k)p^k(1 - p) - p^k\}$$

If we let L_n denote the largest number of consecutive heads in the n flips, then, because L_n will be less than k if (and only if) there are no head strings of length k, the preceding equation can be written as

$$P\{L_n < k\} \approx \exp\{-(n - k)p^k(1 - p) - p^k\}$$

Now, let us suppose that the coin being flipped is fair; that is, suppose that $p = 1/2$. Then the preceding gives

$$P\{L_n < k\} \approx \exp\left\{-\frac{n - k + 2}{2^{k+1}}\right\} \approx \exp\left\{-\frac{n}{2^{k+1}}\right\}$$

where the final approximation supposes that $e^{\frac{k-2}{2^{k+1}}} \approx 1$ (that is, that $\frac{k-2}{2^{k+1}} \approx 0$). Let $j = \log_2 n$, and assume that j is an integer. For $k = j + i$,

$$\frac{n}{2^{k+1}} = \frac{n}{2^j 2^{i+1}} = \frac{1}{2^{i+1}}$$

Consequently,

$$P\{L_n < j + i\} \approx \exp\{-(1/2)^{i+1}\}$$

which implies that

$$\begin{aligned} P\{L_n = j + i\} &= P\{L_n < j + i + 1\} - P\{L_n < j + i\} \\ &\approx \exp\{-(1/2)^{i+2}\} - \exp\{-(1/2)^{i+1}\} \end{aligned}$$

For instance,

$$
\begin{aligned}
P\{L_n < j - 3\} &\approx e^{-4} \approx .0183\\
P\{L_n = j - 3\} &\approx e^{-2} - e^{-4} \approx .1170\\
P\{L_n = j - 2\} &\approx e^{-1} - e^{-2} \approx .2325\\
P\{L_n = j - 1\} &\approx e^{-1/2} - e^{-1} \approx .2387\\
P\{L_n = j\} &\approx e^{-1/4} - e^{-1/2} \approx .1723\\
P\{L_n = j + 1\} &\approx e^{-1/8} - e^{-1/4} \approx .1037\\
P\{L_n = j + 2\} &\approx e^{-1/16} - e^{-1/8} \approx .0569\\
P\{L_n = j + 3\} &\approx e^{-1/32} - e^{-1/16} \approx .0298\\
P\{L_n \geq j + 4\} &\approx 1 - e^{-1/32} \approx .0308
\end{aligned}
$$

Thus, we observe the rather interesting fact that no matter how large n is, the length of the longest run of heads in a sequence of n flips of a fair coin will be within 2 of $\log_2(n) - 1$ with a probability approximately equal to .86.

We now derive an exact expression for the probability that there is a string of k consecutive heads when a coin that lands on heads with probability p is flipped n times. With the events $E_i, i = 1, \ldots, n - k + 1$, as defined earlier, and with L_n denoting, as before, the length of the longest run of heads,

$$P(L_n \geq k) = P(\text{there is a string of } k \text{ consecutive heads}) = P(\cup_{i=1}^{n-k+1} E_i)$$

The inclusion–exclusion identity for the probability of a union can be written as

$$P(\cup_{i=1}^{n-k+1} E_i) = \sum_{r=1}^{n-k+1} (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(E_{i_1} \cdots E_{i_r})$$

Let S_i denote the set of flip numbers to which the event E_i refers. (So, for instance, $S_1 = \{1, \ldots, k + 1\}$.) Now, consider one of the r-way intersection probabilities that does not include the event E_{n-k+1}. That is, consider $P(E_{i_1} \cdots E_{i_r})$ where $i_1 < \cdots < i_r < n - k + 1$. On the one hand, if there is any overlap in the sets $S_{i_1}, \ldots, S_{i_r}$ then this probability is 0. On the other hand, if there is no overlap, then the events $E_{i_1}, \ldots, E_{i_r}$ are independent. Therefore,

$$P(E_{i_1} \cdots E_{i_r}) = \begin{cases} 0, & \text{if there is any overlap in } S_{i_1}, \ldots, S_{i_r} \\ p^{rk}(1 - p)^r, & \text{if there is no overlap} \end{cases}$$

We must now determine the number of different choices of $i_1 < \cdots < i_r < n - k + 1$ for which there is no overlap in the sets $S_{i_1}, \ldots, S_{i_r}$. To do so, note first that each of the $S_{i_j}, j = 1, \ldots, r$, refer to $k + 1$ flips, so, without any overlap, they together refer to $r(k + 1)$ flips. Now consider any permutation of r identical letters a and of $n - r(k + 1)$ identical letters b. Interpret the number of b's before the first a as the number of flips before S_{i_1}, the number of b's between the first and second a as the number of flips between S_{i_1} and S_{i_2}, and so on, with the number of b's after the final a representing the number of flips after S_{i_r}. Because there are $\binom{n-rk}{r}$ permutations of r letters a and of $n - r(k + 1)$ letters b, with every such permutation corresponding (in a one-to-one fashion) to a different nonoverlapping choice, it follows that

$$\sum_{i_1 < \cdots < i_r < n-k+1} P(E_{i_1} \cdots E_{i_r}) = \binom{n - rk}{r} p^{rk}(1 - p)^r$$

We must now consider r-way intersection probabilities of the form

$$P(E_{i_1} \cdots E_{i_{r-1}} E_{n-k+1}),$$

where $i_1 < \cdots < i_{r-1} < n - k + 1$. Now, this probability will equal 0 if there is any overlap in $S_{i_1}, \ldots, S_{i_{r-1}}, S_{n-k}$; if there is no overlap, then the events of the intersection will be independent, so

$$P(E_{i_1} \cdots E_{i_{r-1}} E_{n-k+1}) = [p^k(1 - p)]^{r-1} p^k = p^{kr}(1 - p)^{r-1}$$

By a similar argument as before, the number of nonoverlapping sets $S_{i_1}, \ldots, S_{i_{r-1}}, S_{n-k}$ will equal the number of permutations of $r - 1$ letters a (one for each of the sets $S_{i_1}, \ldots, S_{i_{r-1}}$) and of $n - (r - 1)(k + 1) - k = n - rk - (r - 1)$ letters b (one for each of the trials that are not part of any of $S_{i_1}, \ldots, S_{i_{r-1}}, S_{n-k+1}$). Since there are $\binom{n-rk}{r-1}$ permutations of $r - 1$ letters a and of $n - rk - (r - 1)$ letters b, we have

$$\sum_{i_1 < \ldots < i_{r-1} < n-k+1} P(E_{i_1} \cdots E_{i_{r-1}} E_{n-k+1}) = \binom{n - rk}{r - 1} p^{kr}(1 - p)^{r-1}$$

Putting it all together yields the exact expression, namely,

$$P(L_n \geq k) = \sum_{r=1}^{n-k+1} (-1)^{r+1} \left[\binom{n - rk}{r} + \frac{1}{p} \binom{n - rk}{r - 1} \right] p^{kr}(1 - p)^r$$

where we utilize the convention that $\binom{m}{j} = 0$ if $m < j$.

From a computational point of view, a more efficient method for computing the desired probability than the use of the preceding identity is to derive a set of recursive equations. To do so, let A_n be the event that there is a string of k consecutive heads in a sequence of n flips, and let $P_n = P(A_n)$. We will derive a set of recursive equations for P_n by conditioning on when the first tail appears. For $j = 1, \ldots, k$, let F_j be the event that the first tail appears on flip j, and let H be the event that the first k flips are all heads. Because the events $F_1, \ldots, F_k, H$ are mutually exclusive and exhaustive (that is, exactly one of these events must occur), we have

$$P(A_n) = \sum_{j=1}^{k} P(A_n|F_j)P(F_j) + P(A_n|H)P(H)$$

Now, given that the first tail appears on flip j, where $j < k$, it follows that those j flips are wasted as far as obtaining a string of k heads in a row; thus, the conditional probability of this event is the probability that such a string will occur among the remaining $n - j$ flips. Therefore,

$$P(A_n|F_j) = P_{n-j}$$

Because $P(A_n|H) = 1$, the preceding equation gives

$$\begin{aligned} P_n &= P(A_n) \\ &= \sum_{j=1}^{k} P_{n-j} P(F_j) + P(H) \\ &= \sum_{j=1}^{k} P_{n-j} p^{j-1}(1 - p) + p^k \end{aligned}$$

Starting with $P_j = 0, j < k$, and $P_k = p^k$, we can use the latter formula to recursively compute P_{k+1}, P_{k+2}, and so on, up to P_n. For instance, suppose we want the probability that there is a run of 2 consecutive heads when a fair coin is flipped 4 times. Then, with $k = 2$, we have $P_1 = 0, P_2 = (1/2)^2$. Because, when $p = 1/2$, the recursion becomes

$$P_n = \sum_{j=1}^{k} P_{n-j}(1/2)^j + (1/2)^k$$

we obtain

$$P_3 = P_2(1/2) + P_1(1/2)^2 + (1/2)^2 = 3/8$$

and

$$P_4 = P_3(1/2) + P_2(1/2)^2 + (1/2)^2 = 1/2$$

which is clearly true because there are 8 outcomes that result in a string of 2 consecutive heads: *hhhh*, *hhht*, *hhth*, *hthh*, *thhh*, *hhtt*, *thht*, and *tthh*. Each of these outcomes occurs with probability 1/16. ■

Another use of the Poisson probability distribution arises in situations where "events" occur at certain points in time. One example is to designate the occurrence of an earthquake as an event; another possibility would be for events to correspond to people entering a particular establishment (bank, post office, gas station, and so on); and a third possibility is for an event to occur whenever a war starts. Let us suppose that events are indeed occurring at certain (random) points of time, and let us assume that for some positive constant λ, the following assumptions hold true:

1. The probability that exactly 1 event occurs in a given interval of length h is equal to $\lambda h + o(h)$, where $o(h)$ stands for any function $f(h)$ for which $\lim_{h \to 0} f(h)/h = 0$. [For instance, $f(h) = h^2$ is $o(h)$, whereas $f(h) = h$ is not.]
2. The probability that 2 or more events occur in an interval of length h is equal to $o(h)$.
3. For any integers $n, j_1, j_2, \ldots, j_n$ and any set of n nonoverlapping intervals, if we define E_i to be the event that exactly j_i of the events under consideration occur in the ith of these intervals, then events $E_1, E_2, \ldots, E_n$ are independent.

Loosely put, assumptions 1 and 2 state that for small values of h, the probability that exactly 1 event occurs in an interval of size h equals λh plus something that is small compared with h, whereas the probability that 2 or more events occur is small compared with h. Assumption 3 states that whatever occurs in one interval has no (probability) effect on what will occur in other, nonoverlapping intervals.

We now show that under assumptions 1, 2, and 3, the number of events occurring in any interval of length t is a Poisson random variable with parameter λt. To be precise, let us call the interval $[0, t]$ and denote the number of events occurring in that interval by $N(t)$. To obtain an expression for $P\{N(t) = k\}$, we start by breaking the interval $[0, t]$ into n nonoverlapping subintervals, each of length t/n (Figure 4.9).

Figure 4.9

Now,

$$\begin{aligned} P\{N(t) = k\} = P\{&k \text{ of the } n \text{ subintervals contain exactly 1 event} \\ &\text{and the other } n - k \text{ contain 0 events}\} \\ + P\{&N(t) = k \text{ and at least 1 subinterval contains} \\ &2 \text{ or more events}\} \end{aligned} \tag{7.2}$$

The preceding equation holds because the event on the left side of Equation (7.2), that is, $\{N(t) = k\}$, is clearly equal to the union of the two mutually exclusive events on the right side of the equation. Letting A and B denote the two mutually exclusive events on the right side of Equation (7.2), we have

$$\begin{aligned} P(B) &\leq P\{\text{at least one subinterval contains 2 or more events}\} \\ &= P\left(\bigcup_{i=1}^{n} \{i\text{th subinterval contains 2 or more events}\}\right) \\ &\leq \sum_{i=1}^{n} P\{i\text{th subinterval contains 2 or more events}\} \quad \text{by Boole's inequality} \\ &= \sum_{i=1}^{n} o\left(\frac{t}{n}\right) \quad \text{by assumption 2} \\ &= no\left(\frac{t}{n}\right) \\ &= t\left[\frac{o(t/n)}{t/n}\right] \end{aligned}$$

Now, for any $t, t/n \to 0$ as $n \to \infty$, so $o(t/n)/(t/n) \to 0$ as $n \to \infty$, by the definition of $o(h)$. Hence,

$$P(B) \to 0 \quad \text{as} \quad n \to \infty \tag{7.3}$$

Moreover, since assumptions 1 and 2 imply that[†]

$$\begin{aligned} &P\{0 \text{ events occur in an interval of length } h\} \\ &\quad = 1 - [\lambda h + o(h) + o(h)] = 1 - \lambda h - o(h) \end{aligned}$$

we see from the independence assumption (number 3) that

$$\begin{aligned} P(A) &= P\{k \text{ of the subintervals contain exactly 1 event and the other} \\ &\qquad n - k \text{ contain 0 events}\} \\ &= \binom{n}{k}\left[\frac{\lambda t}{n} + o\left(\frac{t}{n}\right)\right]^k \left[1 - \left(\frac{\lambda t}{n}\right) - o\left(\frac{t}{n}\right)\right]^{n-k} \end{aligned}$$

However, since

$$n\left[\frac{\lambda t}{n} + o\left(\frac{t}{n}\right)\right] = \lambda t + t\left[\frac{o(t/n)}{t/n}\right] \to \lambda t \quad \text{as} \quad n \to \infty$$

[†]The sum of two functions, both of which are $o(h)$, is also $o(h)$. This is so because if $\lim_{h\to 0} f(h)/h = \lim_{h\to 0} g(h)/h = 0$, then $\lim_{h\to 0}[f(h) + g(h)]/h = 0$.

it follows, by the same argument that verified the Poisson approximation to the binomial, that

$$P(A) \to e^{-\lambda t} \frac{(\lambda t)^k}{k!} \quad \text{as} \quad n \to \infty \tag{7.4}$$

Thus, from Equations (7.2), (7.3), and (7.4), by letting $n \to \infty$, we obtain

$$P\{N(t) = k\} = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \quad k = 0, 1, \ldots \tag{7.5}$$

Hence, if assumptions 1, 2, and 3 are satisfied, then the number of events occurring in any fixed interval of length t is a Poisson random variable with mean λt, and we say that the events occur in accordance with a Poisson process having rate λ. The value λ, which can be shown to equal the rate per unit time at which events occur, is a constant that must be empirically determined.

The preceding discussion explains why a Poisson random variable is usually a good approximation for such diverse phenomena as the following:

1. The number of earthquakes occurring during some fixed time span
2. The number of wars per year
3. The number of electrons emitted from a heated cathode during a fixed time period
4. The number of deaths, in a given period of time, of the policyholders of a life insurance company

Example 7e

Suppose that earthquakes occur in the western portion of the United States in accordance with assumptions 1, 2, and 3, with $\lambda = 2$ and with 1 week as the unit of time. (That is, earthquakes occur in accordance with the three assumptions at a rate of 2 per week.)

(a) Find the probability that at least 3 earthquakes occur during the next 2 weeks.

(b) Find the probability distribution of the time, starting from now, until the next earthquake.

Solution (a) From Equation (7.5), we have

$$\begin{aligned} P\{N(2) \geq 3\} &= 1 - P\{N(2) = 0\} - P\{N(2) = 1\} - P\{N(2) = 2\} \\ &= 1 - e^{-4} - 4e^{-4} - \frac{4^2}{2} e^{-4} \\ &= 1 - 13e^{-4} \end{aligned}$$

(b) Let X denote the amount of time (in weeks) until the next earthquake. Because X will be greater than t if and only if no events occur within the next t units of time, we have, from Equation (7.5),

$$P\{X > t\} = P\{N(t) = 0\} = e^{-\lambda t}$$

so the probability distribution function F of the random variable X is given by

$$\begin{aligned} F(t) = P\{X \leq t\} = 1 - P\{X > t\} &= 1 - e^{-\lambda t} \\ &= 1 - e^{-2t} \end{aligned}$$

■

4.7.1 Computing the Poisson Distribution Function

If X is Poisson with parameter λ, then

$$\frac{P\{X = i + 1\}}{P\{X = i\}} = \frac{e^{-\lambda}\lambda^{i+1}/(i + 1)!}{e^{-\lambda}\lambda^{i}/i!} = \frac{\lambda}{i + 1} \tag{7.6}$$

Starting with $P\{X = 0\} = e^{-\lambda}$, we can use (7.6) to compute successively

$$P\{X = 1\} = \lambda P\{X = 0\}$$
$$P\{X = 2\} = \frac{\lambda}{2} P\{X = 1\}$$
$$\vdots$$
$$P\{X = i + 1\} = \frac{\lambda}{i + 1} P\{X = i\}$$

We can use a module to compute the Poisson probabilities for Equation (7.6).

Example 7f

(a) Determine $P\{X \leq 90\}$ when X is Poisson with mean 100.
(b) Determine $P\{Y \leq 1075\}$ when Y is Poisson with mean 1000.

Solution Using the Poisson calculator of StatCrunch yields the solutions:

(a) $P\{X \leq 90\} = .17138$
(b) $P\{Y \leq 1075\} = .99095$ ■

4.8 Other Discrete Probability Distributions

4.8.1 The Geometric Random Variable

Suppose that independent trials, each having a probability $p, 0 < p < 1$, of being a success, are performed until a success occurs. If we let X equal the number of trials required, then

$$P\{X = n\} = (1 - p)^{n-1} p \qquad n = 1, 2, \ldots \tag{8.1}$$

Equation (8.1) follows because, in order for X to equal n, it is necessary and sufficient that the first $n - 1$ trials are failures and the nth trial is a success. Equation (8.1) then follows, since the outcomes of the successive trials are assumed to be independent.

Since

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

it follows that with probability 1, a success will eventually occur. Any random variable X whose probability mass function is given by Equation (8.1) is said to be a *geometric* random variable with parameter p.

Example 8a

An urn contains N white and M black balls. Balls are randomly selected, one at a time, until a black one is obtained. If we assume that each ball selected is replaced before the next one is drawn, what is the probability that

(a) exactly n draws are needed?

(b) at least k draws are needed?

Solution If we let X denote the number of draws needed to select a black ball, then X satisfies Equation (8.1) with $p = M/(M + N)$. Hence,

(a)

$$P\{X = n\} = \left(\frac{N}{M + N}\right)^{n-1} \frac{M}{M + N} = \frac{MN^{n-1}}{(M + N)^n}$$

(b)

$$\begin{aligned} P\{X \geq k\} &= \frac{M}{M + N} \sum_{n=k}^{\infty} \left(\frac{N}{M + N}\right)^{n-1} \\ &= \left(\frac{M}{M + N}\right)\left(\frac{N}{M + N}\right)^{k-1} \Big/ \left[1 - \frac{N}{M + N}\right] \\ &= \left(\frac{N}{M + N}\right)^{k-1} \end{aligned}$$

Of course, part (b) could have been obtained directly, since the probability that at least k trials are necessary to obtain a success is equal to the probability that the first $k - 1$ trials are all failures. That is, for a geometric random variable,

$$P\{X \geq k\} = (1 - p)^{k-1}$$ ■

Example 8b

Find the expected value of a geometric random variable.

Solution With $q = 1 - p$, we have

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} iq^{i-1}p \\ &= \sum_{i=1}^{\infty} (i - 1 + 1)q^{i-1}p \\ &= \sum_{i=1}^{\infty} (i - 1)q^{i-1}p + \sum_{i=1}^{\infty} q^{i-1}p \\ &= \sum_{j=0}^{\infty} jq^j p + 1 \\ &= q \sum_{j=1}^{\infty} jq^{j-1}p + 1 \\ &= qE[X] + 1 \end{aligned}$$

Hence,

$$pE[X] = 1$$

Chapter

5 CONTINUOUS RANDOM VARIABLES

Contents

5.1 Introduction

In Chapter 4, we considered discrete random variables—that is, random variables whose set of possible values is either finite or countably infinite. However, there also exist random variables whose set of possible values is uncountable. Two examples are the time that a train arrives at a specified stop and the lifetime of a transistor. Let X be such a random variable. We say that X is a *continuous* [†] random variable if there exists a nonnegative function f, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers, [‡]

$$P\{X \in B\} = \int_B f(x)\,dx \tag{1.1}$$

The function f is called the *probability density function* of the random variable X. (See Figure 5.1.)

In words, Equation (1.1) 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 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. For instance, from Equation (1.1), letting $B = [a, b]$, we obtain

$$P\{a \le X \le b\} = \int_a^b f(x)\,dx \tag{1.2}$$

[†] Sometimes called *absolutely continuous*.

[‡] Actually, for technical reasons, Equation (1.1) is true only for the *measurable* sets B, which, fortunately, include all sets of practical interest.

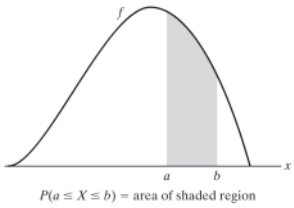


Figure 5.1 Probability density function f.

If we let $a = b$ in Equation (1.2), we get

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

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

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

Example 1a

Suppose that X is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the value of C?
(b) Find $P\{X > 1\}$.

Solution (a) Since f is a probability density function, we must have $\int_{-\infty}^{\infty} f(x)\,dx = 1$, implying that

$$C\int_0^2 (4x - 2x^2)\,dx = 1$$

or

$$C\left[2x^2 - \frac{2x^3}{3}\right]\Bigg|_{x=0}^{x=2} = 1$$

or

$$C = \frac{3}{8}$$

Hence,

(b) $P\{X > 1\} = \int_1^{\infty} f(x)\,dx = \frac{3}{8}\int_1^2 (4x - 2x^2)\,dx = \frac{1}{2}$ ■

Example 1b

The amount of time in hours that a computer functions before breaking down is a continuous random variable with probability density function given by

$$f(x) = \begin{cases} \lambda e^{-x/100} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

What is the probability that

(a) a computer will function between 50 and 150 hours before breaking down?

(b) it will function for fewer than 100 hours?

Solution (a) Since

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \lambda \int_{0}^{\infty} e^{-x/100}\,dx$$

we obtain

$$1 = -\lambda(100)e^{-x/100}\Big|_0^{\infty} = 100\lambda \quad \text{or} \quad \lambda = \frac{1}{100}$$

Hence, the probability that a computer will function between 50 and 150 hours before breaking down is given by

$$\begin{aligned} P\{50 < X < 150\} &= \int_{50}^{150} \frac{1}{100} e^{-x/100}\,dx = -e^{-x/100}\Big|_{50}^{150} \\ &= e^{-1/2} - e^{-3/2} \approx .383 \end{aligned}$$

(b) Similarly,

$$P\{X < 100\} = \int_{0}^{100} \frac{1}{100} e^{-x/100}\,dx = -e^{-x/100}\Big|_0^{100} = 1 - e^{-1} \approx .632$$

In other words, approximately 63.2 percent of the time, a computer will fail before registering 100 hours of use. ■

Example 1c

The lifetime in hours of a certain kind of radio tube is a random variable having a probability density function given by

$$f(x) = \begin{cases} 0 & x \le 100 \\ \dfrac{100}{x^2} & x > 100 \end{cases}$$

What is the probability that exactly 2 of 5 such tubes in a radio set will have to be replaced within the first 150 hours of operation? Assume that the events $E_i, i = 1, 2, 3, 4, 5$, that the ith such tube will have to be replaced within this time are independent.

Solution From the statement of the problem, we have

$$\begin{aligned} P(E_i) &= \int_{0}^{150} f(x)\,dx \\ &= 100 \int_{100}^{150} x^{-2}\,dx \\ &= \frac{1}{3} \end{aligned}$$

Hence, from the independence of the events E_i, it follows that the desired probability is

$$\binom{5}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^3 = \frac{80}{243}$$ ■

The relationship between the cumulative distribution F and the probability density f is expressed by

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

Differentiating both sides of the preceding equation 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 (1.2) as follows:

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

when ε is small and when $f(\cdot)$ is continuous at $x = a$. 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 result, we see that $f(a)$ is a measure of how likely it is that the random variable will be near a.

Example 1d If X is continuous with distribution function F_X and density function f_X, find the density function of $Y = 2X$.

Solution We will determine f_Y in two ways. The first way is to derive, and then differentiate, the distribution function of Y:

$$\begin{aligned} F_Y(a) &= P\{Y \le a\} \\ &= P\{2X \le a\} \\ &= P\{X \le a/2\} \\ &= F_X(a/2) \end{aligned}$$

Differentiation gives

$$f_Y(a) = \frac{1}{2}f_X(a/2)$$

Another way to determine f_Y is to note that

$$\begin{aligned} \epsilon f_Y(a) &\approx P\left\{a - \frac{\epsilon}{2} \le Y \le a + \frac{\epsilon}{2}\right\} \\ &= P\left\{a - \frac{\epsilon}{2} \le 2X \le a + \frac{\epsilon}{2}\right\} \\ &= P\left\{\frac{a}{2} - \frac{\epsilon}{4} \le X \le \frac{a}{2} + \frac{\epsilon}{4}\right\} \\ &\approx \frac{\epsilon}{2}f_X(a/2) \end{aligned}$$

Dividing through by ϵ gives the same result as before. ■

5.2 Expectation and Variance of Continuous Random Variables

In Chapter 4, we defined the expected value of a discrete random variable X by

$$E[X] = \sum_x xP\{X = x\}$$

If X is a continuous random variable having probability density function $f(x)$, then, because

$$f(x)\,dx \approx P\{x \le X \le x + dx\} \quad \text{for } dx \text{ small}$$

it is easy to see that the analogous definition is to define the expected value of X by

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

Example 2a Find $E[X]$ when the density function of X is

$$f(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution

$$\begin{aligned} E[X] &= \int xf(x)\,dx \\ &= \int_0^1 2x^2\,dx \\ &= \frac{2}{3} \end{aligned}$$

■

Example 2b The density function of X is given by

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

Find $E[e^X]$.

Solution Let $Y = e^X$. We start by determining F_Y, the cumulative distribution function of Y. Now, for $1 \le x \le e$,

$$\begin{aligned} F_Y(x) &= P\{Y \le x\} \\ &= P\{e^X \le x\} \\ &= P\{X \le \log(x)\} \\ &= \int_0^{\log(x)} f(y)\,dy \\ &= \log(x) \end{aligned}$$

By differentiating $F_Y(x)$, we can conclude that the probability density function of Y is given by

$$f_Y(x) = \frac{1}{x} \qquad 1 \le x \le e$$

Hence,

$$\begin{aligned} E[e^X] = E[Y] &= \int_{-\infty}^{\infty} x f_Y(x)\,dx \\ &= \int_1^e dx \\ &= e - 1 \end{aligned}$$

■

Although the method employed in Example 2b to compute the expected value of a function of X is always applicable, there is, as in the discrete case, an alternative way of proceeding. The following is a direct analog of Proposition 4.1 of Chapter 4.

Proposition 2.1 If X is a continuous random variable with probability density function $f(x)$, then, for any real-valued function g,

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

An application of Proposition 2.1 to Example 2b yields

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

which is in accord with the result obtained in that example.

The proof of Proposition 2.1 is more involved than that of its discrete random variable analog. We will present such a proof under the provision that the random variable $g(X)$ is nonnegative. (The general proof, which follows the argument in the case we present, is indicated in Theoretical Exercises 5.2 and 5.3.) We will need the following lemma, which is of independent interest.

Lemma 2.1 For a nonnegative random variable Y,

$$E[Y] = \int_0^{\infty} P\{Y > y\}\,dy$$

Proof We present a proof when Y is a continuous random variable with probability density function f_Y. We have

$$\int_0^{\infty} P\{Y > y\}\,dy = \int_0^{\infty} \int_y^{\infty} f_Y(x)\,dx\,dy$$

where we have used the fact that $P\{Y > y\} = \int_y^{\infty} f_Y(x)\,dx$. Interchanging the order of integration in the preceding equation yields

$$\begin{aligned} \int_0^{\infty} P\{Y > y\}\,dy &= \int_0^{\infty} \left(\int_0^x dy \right) f_Y(x)\,dx \\ &= \int_0^{\infty} x f_Y(x)\,dx \\ &= E[Y] \end{aligned}$$

□

Proof of Proposition 2.1 From Lemma 2.1, for any function g for which $g(x) \geq 0$,

$$\begin{aligned} E[g(X)] &= \int_0^\infty P\{g(X) > y\}\, dy \\ &= \int_0^\infty \int_{x:g(x)>y} f(x)\, dx\, dy \\ &= \int_{x:g(x)>0} \int_0^{g(x)} dy\, f(x)\, dx \\ &= \int_{x:g(x)>0} g(x) f(x)\, dx \end{aligned}$$

which completes the proof.

Example 2c A stick of length 1 is split at a point U having density function $f(u) = 1, 0 < u < 1$. Determine the expected length of the piece that contains the point $p, 0 \leq p \leq 1$.

Solution Let $L_p(U)$ denote the length of the substick that contains the point p, and note that

$$L_p(U) = \begin{cases} 1 - U & U < p \\ U & U > p \end{cases}$$

(See Figure 5.2.) Hence, from Proposition 2.1,

$$\begin{aligned} E[L_p(U)] &= \int_0^1 L_p(u)\, du \\ &= \int_0^p (1 - u) du + \int_p^1 u\, du \\ &= \frac{1}{2} - \frac{(1-p)^2}{2} + \frac{1}{2} - \frac{p^2}{2} \\ &= \frac{1}{2} + p(1 - p) \end{aligned}$$

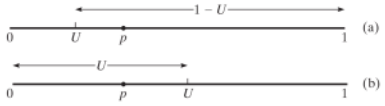


Figure 5.2 Substick containing point p: (a) $U < p$; (b) $U > p$.

Since $p(1 - p)$ is maximized when $p = \frac{1}{2}$, it is interesting to note that the expected length of the substick containing the point p is maximized when p is the midpoint of the original stick. ■

Example 2d Suppose that if you are s minutes early for an appointment, then you incur the cost cs, and if you are s minutes late, then you incur the cost ks. Suppose also that the travel time from where you presently are to the location of your appointment is a continuous random variable having probability density function f. Determine the time at which you should depart if you want to minimize your expected cost.

Solution Let X denote the travel time. If you leave t minutes before your appointment, then your cost—call it $C_t(X)$—is given by

$$C_t(X) = \begin{cases} c(t - X) & \text{if } X \leq t \\ k(X - t) & \text{if } X \geq t \end{cases}$$

Therefore,

$$\begin{aligned} E[C_t(X)] &= \int_0^\infty C_t(x) f(x)\, dx \\ &= \int_0^t c(t - x) f(x)\, dx + \int_t^\infty k(x - t) f(x)\, dx \\ &= ct \int_0^t f(x)\, dx - c \int_0^t x f(x)\, dx + k \int_t^\infty x f(x)\, dx - kt \int_t^\infty f(x)\, dx \end{aligned}$$

The value of t that minimizes $E[C_t(X)]$ can now be obtained by calculus. Differentiation yields

$$\begin{aligned} \frac{d}{dt} E[C_t(X)] &= ctf(t) + cF(t) - ctf(t) - ktf(t) + ktf(t) - k[1 - F(t)] \\ &= (k + c)F(t) - k \end{aligned}$$

Equating the rightmost side to zero shows that the minimal expected cost is obtained when you leave t^* minutes before your appointment, where t^* satisfies

$$F(t^*) = \frac{k}{k + c}$$ ■

As in Chapter 4, we can use Proposition 2.1 to show the following.

Corollary 2.1 If a and b are constants, then

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

The proof of Corollary 2.1 for a continuous random variable X is the same as the one given for a discrete random variable. The only modification is that the sum is replaced by an integral and the probability mass function by a probability density function.

The variance of a continuous random variable is defined exactly as it is for a discrete random variable, namely, if X is a random variable with expected value μ, then the variance of X is defined (for any type of random variable) by

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

The alternative formula,

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

is established in a manner similar to its counterpart in the discrete case.

Example 2e Find $\text{Var}(X)$ for X as given in Example 2a.

Solution We first compute $E[X^2]$.

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)\,dx \\ &= \int_0^1 2x^3\,dx \\ &= \frac{1}{2} \end{aligned}$$

Hence, since $E[X] = \frac{2}{3}$, we obtain

$$\mathrm{Var}(X) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$ ■

It can be shown that, for constants a and b,

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

The proof mimics the one given for discrete random variables.

There are several important classes of continuous random variables that appear frequently in applications of probability; the next few sections are devoted to a study of some of them.

5.3 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

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Note that Equation (3.1) is a density function, since $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 dx = 1$. Because $f(x) > 0$ only when $x \in (0, 1)$, it follows that X must assume a value in interval (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 it is to be near any other value. To verify this statement, note that for any $0 < a < b < 1$,

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

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 the probability density function of X is given by

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

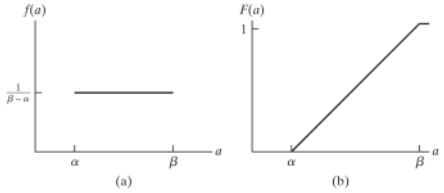


Figure 5.3 Graph of (a) $f(a)$ and (b) $F(a)$ for a uniform (α, β) random variable.

Since $F(a) = \int_{-\infty}^{a} f(x)\,dx$, it follows from Equation (3.2) that the distribution function of a uniform random variable on the interval (α, β) is given by

$$F(a) = \begin{cases} 0 & a \le \alpha \\ \dfrac{a - \alpha}{\beta - \alpha} & \alpha < a < \beta \\ 1 & a \ge \beta \end{cases}$$

Figure 5.3 presents a graph of $f(a)$ and $F(a)$.

Example 3a Let X be uniformly distributed over (α, β). Find (a) $E[X]$ and (b) $\mathrm{Var}(X)$.

Solution (a)

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

In words, the expected value of a random variable that is uniformly distributed over some interval is equal to the midpoint of that interval.

(b) To find $\mathrm{Var}(X)$, we first calculate $E[X^2]$.

$$\begin{aligned} E[X^2] &= \int_{\alpha}^{\beta} \frac{1}{\beta - \alpha} x^2\,dx \\ &= \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} \\ &= \frac{\beta^2 + \alpha\beta + \alpha^2}{3} \end{aligned}$$

Hence,

$$\begin{aligned} \mathrm{Var}(X) &= \frac{\beta^2 + \alpha\beta + \alpha^2}{3} - \frac{(\alpha + \beta)^2}{4} \\ &= \frac{(\beta - \alpha)^2}{12} \end{aligned}$$

Therefore, the variance of a random variable that is uniformly distributed over some interval is the square of the length of that interval divided by 12. ■

Example 3b

If X is uniformly distributed over (0, 10), calculate the probability that (a) $X < 3$, (b) $X > 6$, and (c) $3 < X < 8$.

Solution (a) $P\{X < 3\} = \int_0^3 \frac{1}{10}\,dx = \frac{3}{10}$

(b) $P\{X > 6\} = \int_6^{10} \frac{1}{10}\,dx = \frac{4}{10}$

(c) $P\{3 < X < 8\} = \int_3^8 \frac{1}{10}\,dx = \frac{1}{2}$ ■

Example 3c

Buses arrive at a specified stop at 15-minute intervals starting at 7 A.M. That is, they arrive at 7, 7:15, 7:30, 7:45, and so on. If a passenger arrives at the stop at a time that is uniformly distributed between 7 and 7:30, find the probability that he waits

(a) less than 5 minutes for a bus;

(b) more than 10 minutes for a bus.

Solution Let X denote the number of minutes past 7 that the passenger arrives at the stop. Since X is a uniform random variable over the interval (0, 30), it follows that the passenger will have to wait less than 5 minutes if (and only if) he arrives between 7:10 and 7:15 or between 7:25 and 7:30. Hence, the desired probability for part (a) is

$$P\{10 < X < 15\} + P\{25 < X < 30\} = \int_{10}^{15} \frac{1}{30}\,dx + \int_{25}^{30} \frac{1}{30}\,dx = \frac{1}{3}$$

Similarly, he would have to wait more than 10 minutes if he arrives between 7 and 7:05 or between 7:15 and 7:20, so the probability for part (b) is

$$P\{0 < X < 5\} + P\{15 < X < 20\} = \frac{1}{3}$$ ■

The next example was first considered by the French mathematician Joseph L. F. Bertrand in 1889 and is often referred to as *Bertrand's paradox*. It represents our initial introduction to a subject commonly referred to as *geometrical probability*.

Example 3d

Consider a random chord of a circle. What is the probability that the length of the chord will be greater than the side of the equilateral triangle inscribed in that circle?

Solution As stated, the problem is incapable of solution because it is not clear what is meant by a random chord. To give meaning to this phrase, we shall reformulate the problem in two distinct ways.

The first formulation is as follows: The position of the chord can be determined by its distance from the center of the circle. This distance can vary between 0 and r, the radius of the circle. Now, the length of the chord will be greater than the side of the equilateral triangle inscribed in the circle if the distance from the chord to the center of the circle is less than $r/2$. Hence, by assuming that a random chord is a chord whose distance D from the center of the circle is uniformly distributed between 0 and r, we see that the probability that the length of the chord is greater than the side of an inscribed equilateral triangle is

$$P\left\{D < \frac{r}{2}\right\} = \frac{r/2}{r} = \frac{1}{2}$$

Chapter

6

JOINTLY DISTRIBUTED RANDOM VARIABLES

Contents

6.1 Joint Distribution Functions

Thus far, we have concerned ourselves only with probability distributions for single random variables. However, we are often interested in probability statements concerning two or more random variables. In order 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

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

All joint probability statements about X and Y can, in theory, be answered in terms of their joint distribution function. For instance,

$$P(a_1 < X \le a_2, b_1 < Y \le b_2) = F(a_2,b_2) + F(a_1,b_1) - F(a_1,b_2) - F(a_2,b_1) \quad (1.1)$$

whenever $a_1 < a_2$, $b_1 < b_2$. To verify Equation (1.1), note that for $a_1 < a_2$,

$$P(X \le a_2, Y \le b) = P(X \le a_1, Y \le b) + P(a_1 < X \le a_2, Y \le b)$$

giving that

$$P(a_1 < X \le a_2, Y \le b) = F(a_2,b) - F(a_1,b) \quad (1.2)$$

Also, because for $b_1 < b_2$,

$$P(a_1 < X \le a_2, Y \le b_2) = P(a_1 < X \le a_2, Y \le b_1) + P(a_1 < X \le a_2, b_1 < Y \le b_2)$$

we have that when $a_1 < a_2, b_1 < b_2$

$$\begin{aligned} P(a_1 < X \le a_2, b_1 < Y \le b_2) &= P(a_1 < X \le a_2, Y \le b_2) \\ &\quad - P(a_1 < X \le a_2, Y \le b_1) \\ &= F(a_2, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, b_1) \end{aligned}$$

where the final equality used Equation (1.2).
When X and Y are discrete random variables, with X taking on one of the values x_i, $i \ge 1$, and Y one of the values y_j, $j \ge 1$, it is convenient to define the *joint probability mass function* of X and Y by

$$p(x, y) = P(X = x, Y = y)$$

Using that the event $\{X = x\}$ is the union of the mutually exclusive events $\{X = x, Y = y_j\}$, $j \ge 1$, it follows that the probability mass function of X can be obtained from the joint probability mass function by

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= P(\cup_j \{X = x, Y = y_j\}) \\ &= \sum_j P(X = x, Y = y_j) \\ &= \sum_j p(x, y_j) \end{aligned}$$

Similarly, the probability mass function of Y is obtained from

$$p_Y(y) = \sum_i p(x_i, y)$$

Example 1a

Suppose that 3 balls are randomly selected from an urn containing 3 red, 4 white, and 5 blue balls. If we let X and Y denote, respectively, the number of red and white balls chosen, then the joint probability mass function of X and Y, $p(i, j) = P\{X = i, Y = j\}$, is obtained by noting that $X = i$, $Y = j$ if, of the 3 balls selected, i are red, j are white, and $3 - i - j$ are blue. Because all subsets of size 3 are equally likely to be chosen, it follows that

$$p(i, j) = \frac{\binom{3}{i}\binom{4}{j}\binom{5}{3-i-j}}{\binom{12}{3}}$$

Consequently,

$$\begin{aligned} p(0, 0) &= \binom{5}{3} \Big/ \binom{12}{3} = \frac{10}{220} \\ p(0, 1) &= \binom{4}{1}\binom{5}{2} \Big/ \binom{12}{3} = \frac{40}{220} \\ p(0, 2) &= \binom{4}{2}\binom{5}{1} \Big/ \binom{12}{3} = \frac{30}{220} \end{aligned}$$

$$p(0,3) = \binom{4}{3}\Big/\binom{12}{3} = \frac{4}{220}$$

$$p(1,0) = \binom{3}{1}\binom{5}{2}\Big/\binom{12}{3} = \frac{30}{220}$$

$$p(1,1) = \binom{3}{1}\binom{4}{1}\binom{5}{1}\Big/\binom{12}{3} = \frac{60}{220}$$

$$p(1,2) = \binom{3}{1}\binom{4}{2}\Big/\binom{12}{3} = \frac{18}{220}$$

$$p(2,0) = \binom{3}{2}\binom{5}{1}\Big/\binom{12}{3} = \frac{15}{220}$$

$$p(2,1) = \binom{3}{2}\binom{4}{1}\Big/\binom{12}{3} = \frac{12}{220}$$

$$p(3,0) = \binom{3}{3}\Big/\binom{12}{3} = \frac{1}{220}$$

These probabilities can most easily be expressed in tabular form, as in Table 6.1. The reader should note that the probability mass function of X is obtained by computing the row sums, whereas the probability mass function of Y is obtained by computing the column sums. Because the individual probability mass functions of X and Y thus appear in the margin of such a table, they are often referred to as the *marginal probability mass functions* of X and Y, respectively. ■

Table 6.1 $P\{X = i, Y = j\}$.

i \ j	0	1	2	3	Row sum $= P\{X = i\}$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
Column sum $= P\{Y = j\}$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

Example 1b

Suppose that 15 percent of the families in a certain community have no children, 20 percent have 1 child, 35 percent have 2 children, and 30 percent have 3. Suppose further that in each family each child is equally likely (independently) to be a boy or a girl. If a family is chosen at random from this community, then B, the number of boys, and G, the number of girls, in this family will have the joint probability mass function shown in Table 6.2.

Table 6.2 $P\{B = i, G = j\}$.

i \ j	0	1	2	3	Row sum $= P\{B = i\}$
0	.15	.10	.0875	.0375	.3750
1	.10	.175	.1125	0	.3875
2	.0875	.1125	0	0	.2000
3	.0375	0	0	0	.0375
Column sum $= P\{G = j\}$	.3750	.3875	.2000	.0375	

The probabilities shown in Table 6.2 are obtained as follows:

$$P\{B = 0, G = 0\} = P\{\text{no children}\} = .15$$

$$P\{B = 0, G = 1\} = P\{1 \text{ girl and total of 1 child}\}$$
$$= P\{1 \text{ child}\}P\{1 \text{ girl}|1 \text{ child}\} = (.20)\left(\frac{1}{2}\right)$$

$$P\{B = 0, G = 2\} = P\{2 \text{ girls and total of 2 children}\}$$
$$= P\{2 \text{ children}\}P\{2 \text{ girls}|2 \text{ children}\} = (.35)\left(\frac{1}{2}\right)^2$$

We leave the verification of the remaining probabilities in the table to the reader. ■

Example 1c

Consider independent trials where each trial is a success with probability p. Let X_r denote the number of trials until there have been r successes, and let Y_s denote the number of trials until there have been s failures. Suppose we want to derive their joint probability mass function $P(X_r = i, Y_s = j)$. To do so, first consider the case $i < j$. In this case, write

$$P(X_r = i, Y_s = j) = P(X_r = i)P(Y_s = j|X_r = i)$$

Now, if there have been r successes after trial i then there have been $i - r$ failures by that point. Hence, the conditional distribution of Y_s, given that $X_r = i$, is the distribution of i plus the number of additional trials after trial i until there have been an additional $s - i + r$ failures. Hence,

$$P(X_r = i, Y_s = j) = P(X_r = i)P(Y_{s-i+r} = j - i), \quad i < j$$

Because X_r is a negative binomial random variable with parameters (r, p) and Y_{s-i+r} is a negative binomial random variable with parameters $(s - i + r, 1 - p)$, the preceding yields

$$P(X_r = i, Y_s = j) = \binom{i-1}{r-1}p^r(1-p)^{i-r}\binom{j-i-1}{s-i+r-1}(1-p)^{s-i+r}p^{j-s-r}, \quad i < j$$

We leave it as an exercise to determine the analogous expression when $j < i$. ■

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 every set C of pairs of real numbers (that is, C is a set in the two-dimensional plane),

$$P\{(X,Y) \in C\} = \iint_{(x,y)\in C} f(x,y)\,dx\,dy \tag{1.3}$$

The function $f(x,y)$ is called the *joint probability density function* of X and Y. If A and B are any sets of real numbers, then by defining $C = \{(x,y) : x \in A, y \in B\}$, we see from Equation (1.3) that

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x,y)\,dx\,dy \tag{1.4}$$

Because

$$\begin{aligned} F(a,b) &= P\{X \in (-\infty,a], Y \in (-\infty,b]\} \\ &= \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)\,dx\,dy \end{aligned}$$

it follows, upon differentiation, that

$$f(a,b) = \frac{\partial^2}{\partial a \partial b} F(a,b)$$

wherever the partial derivatives are defined. Another interpretation of the joint density function, obtained from Equation (1.4), is

$$\begin{aligned} P\{a < X < a + da, b < Y < b + db\} &= \int_b^{d+db}\int_a^{a+da} f(x,y)\,dx\,dy \\ &\approx f(a,b)\,da\,db \end{aligned}$$

when da and db are small and $f(x,y)$ is continuous at a, b. Hence, $f(a,b)$ is a measure of how likely it is that the random vector (X, Y) will be near (a, b).

If X and Y are jointly continuous, they are individually continuous, and their probability density functions can be obtained as follows:

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

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

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

Example 1d

The joint density function of X and Y is given by

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty,\ 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Compute (a) $P\{X > 1, Y < 1\}$, (b) $P\{X < Y\}$, and (c) $P\{X < a\}$.

Solution

(a)
$$P(X > 1, Y < 1) = \int_0^1 \int_1^{\infty} 2e^{-x}e^{-2y}\,dx\,dy$$

Now,

$$\int_1^{\infty} e^{-x}dx = -e^{-x}\Big|_1^{\infty} = e^{-1}$$

giving that

$$P(X > 1, Y < 1) = e^{-1}\int_0^1 2e^{-2y}dy = e^{-1}(1 - e^{-2})$$

(b)
$$\begin{aligned} P\{X < Y\} &= \iint_{(x,y):x<y} 2e^{-x}e^{-2y}\,dx\,dy \\ &= \int_0^{\infty}\int_0^{y} 2e^{-x}e^{-2y}\,dx\,dy \\ &= \int_0^{\infty} 2e^{-2y}(1 - e^{-y})dy \\ &= \int_0^{\infty} 2e^{-2y}dy - \int_0^{\infty} 2e^{-3y}dy \\ &= 1 - \frac{2}{3} \\ &= \frac{1}{3} \end{aligned}$$

(c)
$$\begin{aligned} P\{X < a\} &= \int_0^a \int_0^{\infty} 2e^{-2y}e^{-x}\,dy\,dx \\ &= \int_0^a e^{-x}\,dx \\ &= 1 - e^{-a} \end{aligned}$$

■

Example 1e

Consider a circle of radius R, and suppose that a point within the circle is randomly chosen in such a manner that all regions within the circle of equal area are equally likely to contain the point. (In other words, the point is uniformly distributed within the circle.) If we let the center of the circle denote the origin and define X and Y to be the coordinates of the point chosen (Figure 6.1), then, since (X, Y) is equally likely to be near each point in the circle, it follows that the joint density function of X and Y is given by

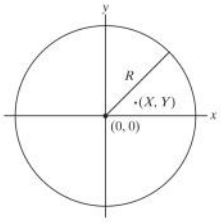


Figure 6.1 Joint probability distribution.

$$f(x,y) = \begin{cases} c & \text{if } x^2 + y^2 \le R^2 \\ 0 & \text{if } x^2 + y^2 > R^2 \end{cases}$$

for some value of c.

(a) Determine c.

(b) Find the marginal density functions of X and Y.

(c) Compute the probability that D, the distance from the origin of the point selected, is less than or equal to a.

(d) Find $E[D]$.

Solution

(a) Because

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

it follows that

$$c \iint_{x^2+y^2\le R^2} dy\,dx = 1$$

We can evaluate $\iint_{x^2+y^2\le R^2} dy\,dx$ either by using polar coordinates or, more simply, by noting that it represents the area of the circle and is thus equal to πR^2. Hence,

$$c = \frac{1}{\pi R^2}$$

(b)
$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \frac{1}{\pi R^2}\int_{x^2+y^2\le R^2} dy \\ &= \frac{1}{\pi R^2}\int_{-a}^{a} dy, \quad \text{where } a = \sqrt{R^2 - x^2} \\ &= \frac{2}{\pi R^2}\sqrt{R^2 - x^2}, \quad x^2 \le R^2 \end{aligned}$$

and it equals 0 when $x^2 > R^2$. By symmetry, the marginal density of Y is given by

$$\begin{aligned} f_Y(y) &= \frac{2}{\pi R^2}\sqrt{R^2 - y^2}, \quad y^2 \le R^2 \\ &= 0 \qquad\qquad\qquad\quad y^2 > R^2 \end{aligned}$$

(c) The distribution function of $D = \sqrt{X^2 + Y^2}$, the distance from the origin, is obtained as follows: For $0 \le a \le R$,

$$\begin{aligned} F_D(a) &= P\{\sqrt{X^2 + Y^2} \le a\} \\ &= P\{X^2 + Y^2 \le a^2\} \\ &= \iint\limits_{x^2+y^2 \le a^2} f(x,y)\,dy\,dx \\ &= \frac{1}{\pi R^2} \iint\limits_{x^2+y^2 \le a^2} dy\,dx \\ &= \frac{\pi a^2}{\pi R^2} \\ &= \frac{a^2}{R^2} \end{aligned}$$

where we have used the fact that $\iint_{x^2+y^2 \le a^2} dy\,dx$ is the area of a circle of radius a and thus is equal to πa^2.

(d) From part (c), the density function of D is

$$f_D(a) = \frac{2a}{R^2} \quad 0 \le a \le R$$

Hence,

$$E[D] = \frac{2}{R^2}\int_0^R a^2 da = \frac{2R}{3}$$ ■

Example If The joint density of X and Y is given by

$$f(x,y) = \begin{cases} e^{-(x+y)} & 0 < x < \infty,\ 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Find the density function of the random variable X/Y.

Solution We start by computing the distribution function of X/Y. For $a > 0$,

$$\begin{aligned} F_{X/Y}(a) &= P\left\{\frac{X}{Y} \le a\right\} \\ &= \iint_{x/y \le a} e^{-(x+y)}\,dx\,dy \\ &= \int_0^\infty \int_0^{ay} e^{-(x+y)}\,dx\,dy \\ &= \int_0^\infty (1 - e^{-ay})e^{-y}\,dy \\ &= \left\{-e^{-y} + \frac{e^{-(a+1)y}}{a+1}\right\}\Bigg|_0^\infty \\ &= 1 - \frac{1}{a+1} \end{aligned}$$

Differentiation shows that the density function of X/Y is given by $f_{X/Y}(a) = 1/(a+1)^2, 0 < a < \infty$. ■

We can also define joint probability distributions for n random variables in exactly the same manner as we did for $n = 2$. For instance, the joint cumulative probability distribution function $F(a_1, a_2, \ldots, a_n)$ of the n random variables $X_1, X_2, \ldots, X_n$ is defined by

$$F(a_1, a_2, \ldots, a_n) = P\{X_1 \le a_1, X_2 \le a_2, \ldots, X_n \le a_n\}$$

Further, the n random variables are said to be *jointly continuous* if there exists a function $f(x_1, x_2, \ldots, x_n)$, called the *joint probability density function*, such that, for any set C in n-space,

$$P\{(X_1, X_2, \ldots, X_n) \in C\} = \iint\cdots\int_{(x_1,\ldots,x_n)\in C} f(x_1, \ldots, x_n)dx_1dx_2\cdots dx_n$$

In particular, for any n sets of real numbers $A_1, A_2, \ldots, A_n$,

$$\begin{aligned} &P\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\} \\ &\quad = \int_{A_n}\int_{A_{n-1}}\cdots\int_{A_1} f(x_1, \ldots, x_n)\,dx_1dx_2\cdots dx_n \end{aligned}$$

Example 1g **The multinomial distribution**

One of the most important joint distributions is the multinomial distribution, which arises when a sequence of n independent and identical experiments is performed. Suppose that each experiment can result in any one of r possible outcomes, with respective probabilities $p_1, p_2, \ldots, p_r$, $\sum_{i=1}^{r} p_i = 1$. If we let X_i denote the number of the n experiments that result in outcome number i, then

$$P\{X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r\} = \frac{n!}{n_1!n_2!\cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r} \tag{1.5}$$

whenever $\sum_{i=1}^{r} n_i = n$.

Equation (1.5) is verified by noting that any sequence of outcomes for the n experiments that leads to outcome i occurring n_i times for $i = 1, 2, \ldots, r$ will, by the assumed independence of experiments, have probability $p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}$ of occurring. Because there are $n!/(n_1!n_2! \ldots n_r!)$ such sequences of outcomes (there are $n!/n_1! \ldots n_r!$ different permutations of n things of which n_1 are alike, n_2 are alike, $\ldots, n_r$ are alike), Equation (1.5) is established. The joint distribution whose joint probability mass function is specified by Equation (1.5) is called the *multinomial distribution*. Note that when $r = 2$, the multinomial reduces to the binomial distribution.

Note also that any sum of a fixed set of the $X_i's$ will have a binomial distribution. That is, if $N \subset \{1, 2, \ldots, r\}$, then $\sum_{i \in N} X_i$ will be a binomial random variable with parameters n and $p = \sum_{i \in N} p_i$. This follows because $\sum_{i \in N} X_i$ represents the number of the n experiments whose outcome is in N, and each experiment will independently have such an outcome with probability $\sum_{i \in N} p_i$.

As an application of the multinomial distribution, suppose that a fair die is rolled 9 times. The probability that 1 appears three times, 2 and 3 twice each, 4 and 5 once each, and 6 not at all is

$$\frac{9!}{3!2!2!1!1!0!} \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 = \frac{9!}{3!2!2!} \left(\frac{1}{6}\right)^9$$

We can also use the multinomial distribution to analyze a variation of the classical birthday problem which asks for the probability that no 3 people in a group of size n have the same birthday when the birthdays of the n people are independent and each birthday is equally likely to be any of the 365 days of the year. Because this probability is 0 when $n > 730$ (why is this), we will suppose that $n \leq 730$. To find the desired probability, note that there will be no set of 3 people having the same birthday if each of the 365 days of the year is the birthday of at most 2 persons. Now, this will be the case if for some $i \leq n/2$ the event A_i occurs, where A_i is the event that the 365 days of the year can be partitioned into three groups of respective sizes $i, n - 2i$, and $365 - n + i$ such that every day in the first group is the birthday of exactly 2 of the n individuals, every day in the second group is the birthday of exactly 1 of the n individuals, and every day in the third group is the birthday of none of the n individuals. Now, because each day of the year is equally likely to be the birthday of an individual, it follows, for a given partition of the 365 days into three groups of respective sizes $i, n - 2i$, and $365 - n + i$, that the probability each day in the first group is the birthday of exactly 2 of the n individuals, each day in the second group is the birthday of exactly 1 of the n individuals, and each day in the third group is the birthday of none of the n individuals is equal to the multinomial probability

$$\frac{n!}{(2!)^i(1!)^{n-2i}(0!)^{365-n+i}}\left(\frac{1}{365}\right)^n.$$

As the number of partitions of the 365 days of the year into 3 groups of respective sizes $i, n-2i, 365-n+i$ is $\frac{365!}{i!(n-2i)!(365-n+i)!}$, it follows that

$$P(A_i) = \frac{365!}{i!(n-2i)!(365-n+i)!}\frac{n!}{2^i}\left(\frac{1}{365}\right)^n, \quad i \le n/2$$

As the events $A_i, i \le n/2$, are mutually exclusive we have that

$$P\{\text{no set of three with same birthday}\} = \sum_{i=0}^{\lfloor n/2 \rfloor} \frac{365!}{i!(n-2i)!(365-n+i)!}\frac{n!}{2^i}\left(\frac{1}{365}\right)^n$$

When $n = 88$, the preceding gives

$$P\{\text{no set of three with same birthday}\} = \sum_{i=0}^{44} \frac{365!}{i!(88-2i)!(277+i)!}\frac{88!}{2^i}\left(\frac{1}{365}\right)^{88} \approx .504$$

■

6.2 Independent Random Variables

The random variables X and Y are said to be *independent* if, for any two sets of real numbers A and B,

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \tag{2.1}$$

In other words, X and Y are independent if, for all A and B, the events $E_A = \{X \in A\}$ and $F_B = \{Y \in B\}$ are independent.

It can be shown by using the three axioms of probability that Equation (2.1) will follow if and only if, for all a, b,

$$P\{X \le a, Y \le b\} = P\{X \le a\}P\{Y \le b\}$$

Hence, in terms of the joint distribution function F of X and Y, X and Y are independent if

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

When X and Y are discrete random variables, the condition of independence (2.1) is equivalent to

$$p(x,y) = p_X(x)p_Y(y) \quad \text{for all } x, y \tag{2.2}$$

The equivalence follows because, if Equation (2.1) is satisfied, then we obtain Equation (2.2) by letting A and B be, respectively, the one-point sets $A = \{x\}$ and $B = \{y\}$. Furthermore, if Equation (2.2) is valid, then for any sets A, B,

$$\begin{aligned} P\{X \in A, Y \in B\} &= \sum_{y \in B}\sum_{x \in A} p(x,y) \\ &= \sum_{y \in B}\sum_{x \in A} p_X(x)p_Y(y) \\ &= \sum_{y \in B} p_Y(y) \sum_{x \in A} p_X(x) \\ &= P\{Y \in B\}P\{X \in A\} \end{aligned}$$

and Equation (2.1) is established.

In the jointly continuous case, the condition of independence is equivalent to

$$f(x, y) = f_X(x) f_Y(y) \quad \text{for all } x, y$$

Thus, loosely speaking, X and Y are independent if knowing the value of one does not change the distribution of the other. Random variables that are not independent are said to be *dependent*.

Example 2a

Suppose that $n + m$ independent trials having a common probability of success p are performed. If X is the number of successes in the first n trials, and Y is the number of successes in the final m trials, then X and Y are independent, since knowing the number of successes in the first n trials does not affect the distribution of the number of successes in the final m trials (by the assumption of independent trials). In fact, for integral x and y,

$$\begin{aligned} P\{X = x, Y = y\} &= \binom{n}{x} p^x (1 - p)^{n-x} \binom{m}{y} p^y (1 - p)^{m-y} \qquad \begin{matrix} 0 \le x \le n, \\ 0 \le y \le m \end{matrix} \\ &= P\{X = x\} P\{Y = y\} \end{aligned}$$

In contrast, X and Z will be dependent, where Z is the total number of successes in the $n + m$ trials. (Why?) ■

Example 2b

Suppose that the number of people who enter a post office on a given day is a Poisson random variable with parameter λ. Show that if each person who enters the post office is a male with probability p and a female with probability $1 - p$, then the number of males and females entering the post office are independent Poisson random variables with respective parameters λp and $\lambda(1 - p)$.

Solution Let X and Y denote, respectively, the number of males and females that enter the post office. We shall show the independence of X and Y by establishing Equation (2.2). To obtain an expression for $P\{X = i, Y = j\}$, we condition on whether or not $X + Y = i + j$. This gives:

$$\begin{aligned} P\{X = i, Y = j\} = {} & P\{X = i, Y = j | X + Y = i + j\} P\{X + Y = i + j\} \\ & + P\{X = i, Y = j | X + Y \neq i + j\} P\{X + Y \neq i + j\} \end{aligned}$$

[Note that this equation is merely a special case of the formula $P(E) = P(E|F)P(F) + P(E|F^c)P(F^c)$.]

Since $P\{X = i, Y = j | X + Y \neq i + j\}$ is clearly 0, we obtain

$$P\{X = i, Y = j\} = P\{X = i, Y = j | X + Y = i + j\} P\{X + Y = i + j\} \tag{2.3}$$

Now, because $X + Y$ is the total number of people who enter the post office, it follows, by assumption, that

$$P\{X + Y = i + j\} = e^{-\lambda} \frac{\lambda^{i+j}}{(i + j)!} \tag{2.4}$$

Furthermore, given that $i + j$ people do enter the post office, since each person entering will be male with probability p, it follows that the probability that exactly i of them will be male (and thus j of them female) is just the binomial probability $\binom{i + j}{i} p^i (1 - p)^j$. That is,

$$P\{X = i, Y = j | X + Y = i + j\} = \binom{i+j}{i} p^i (1-p)^j \tag{2.5}$$

Substituting Equations (2.4) and (2.5) into Equation (2.3) yields

$$\begin{aligned} P\{X = i, Y = j\} &= \binom{i+j}{i} p^i (1-p)^j e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!} \\ &= e^{-\lambda} \frac{(\lambda p)^i}{i!j!} [\lambda(1-p)]^j \\ &= \frac{e^{-\lambda p}(\lambda p)^i}{i!} e^{-\lambda(1-p)} \frac{[\lambda(1-p)]^j}{j!} \end{aligned} \tag{2.6}$$

Hence,

$$P\{X = i\} = e^{-\lambda p} \frac{(\lambda p)^i}{i!} \sum_j e^{-\lambda(1-p)} \frac{[\lambda(1-p)]^j}{j!} = e^{-\lambda p} \frac{(\lambda p)^i}{i!} \tag{2.7}$$

and similarly,

$$P\{Y = j\} = e^{-\lambda(1-p)} \frac{[\lambda(1-p)]^j}{j!} \tag{2.8}$$

Equations (2.6), (2.7), and (2.8) establish the desired result. ■

Example 2c

A man and a woman decide to meet at a certain location. If each of them independently arrives at a time uniformly distributed between 12 noon and 1 P.M., find the probability that the first to arrive has to wait longer than 10 minutes.

Solution If we let X and Y denote, respectively, the time past 12 that the man and the woman arrive, then X and Y are independent random variables, each of which is uniformly distributed over $(0, 60)$. The desired probability, $P\{X + 10 < Y\} + P\{Y + 10 < X\}$, which, by symmetry, equals $2P\{X + 10 < Y\}$, is obtained as follows:

$$\begin{aligned} 2P\{X + 10 < Y\} &= 2 \iint_{x+10<y} f(x,y)\,dx\,dy \\ &= 2 \iint_{x+10<y} f_X(x) f_Y(y)\,dx\,dy \\ &= 2 \int_{10}^{60} \int_{0}^{y-10} \left(\frac{1}{60}\right)^2 dx\,dy \\ &= \frac{2}{(60)^2} \int_{10}^{60} (y - 10)\,dy \\ &= \frac{25}{36} \end{aligned}$$

■

Our next example presents the oldest problem dealing with geometrical probabilities. It was first considered and solved by Buffon, a French naturalist of the eighteenth century, and is usually referred to as *Buffon's needle problem*.

Chapter 7
PROPERTIES OF EXPECTATION

Contents

7.1 Introduction

In this chapter, we develop and exploit additional properties of expected values. To begin, recall that the expected value of the random variable X is defined by

$$E[X] = \sum_x x\,p(x)$$

when X is a discrete random variable with probability mass function $p(x)$, and by

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

when X is a continuous random variable with probability density function $f(x)$.

Since $E[X]$ is a weighted average of the possible values of X, it follows that if X must lie between a and b, then so must its expected value. That is, if

$$P\{a \le X \le b\} = 1$$

then

$$a \le E[X] \le b$$

To verify the preceding statement, suppose that X is a discrete random variable for which $P\{a \le X \le b\} = 1$. Since this implies that $p(x) = 0$ for all x outside of the interval $[a, b]$, it follows that

$$\begin{aligned} E[X] &= \sum_{x:p(x)>0} x\,p(x) \\ &\ge \sum_{x:p(x)>0} a\,p(x) \end{aligned}$$

$$= a \sum_{x:p(x)>0} p(x)$$
$$= a$$

In the same manner, it can be shown that $E[X] \leq b$, so the result follows for discrete random variables. As the proof in the continuous case is similar, the result follows.

7.2 Expectation of Sums of Random Variables

For a two-dimensional analog of Propositions 4.1 of Chapter 4 and 2.1 of Chapter 5, which give the computational formulas for the expected value of a function of a random variable, suppose that X and Y are random variables and g is a function of two variables. Then we have the following result.

Proposition 2.1 If X and Y have a joint probability mass function $p(x,y)$, then

$$E[g(X,Y)] = \sum_y \sum_x g(x,y)\, p(x,y)$$

If X and Y have a joint probability density function $f(x,y)$, then

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

Let us give a proof of Proposition 2.1 when the random variables X and Y are jointly continuous with joint density function $f(x,y)$ and when $g(X,Y)$ is a nonnegative random variable. Because $g(X,Y) \geq 0$, we have, by Lemma 2.1 of Chapter 5, that

$$E[g(X,Y)] = \int_0^{\infty} P\{g(X,Y) > t\}\, dt$$

Writing

$$P\{g(X,Y) > t\} = \int \int_{(x,y):g(x,y)>t} f(x,y)\, dy\, dx$$

shows that

$$E[g(X,Y)] = \int_0^{\infty} \int \int_{(x,y):g(x,y)>t} f(x,y)\, dy\, dx\, dt$$

Interchanging the order of integration gives

$$E[g(X,Y)] = \int_x \int_y \int_{t=0}^{g(x,y)} f(x,y)\, dt\, dy\, dx$$
$$= \int_x \int_y g(x,y)\, f(x,y)\, dy\, dx$$

Thus, the result is proven when $g(X,Y)$ is a nonnegative random variable. The general case then follows as in the one-dimensional case. (See Theoretical Exercises 2 and 3 of Chapter 5.)

Example 2a An accident occurs at a point X that is uniformly distributed on a road of length L. At the time of the accident, an ambulance is at a location Y that is also uniformly distributed on the road. Assuming that X and Y are independent, find the expected distance between the ambulance and the point of the accident.

We note from Equation (6.5) that if ρ is near $+1$ or -1, then the mean square error of the best linear predictor is near zero. ■

Example 6d An example in which the conditional expectation of Y given X is linear in X, and hence in which the best linear predictor of Y with respect to X is the best overall predictor, is when X and Y have a bivariate normal distribution. For, as shown in Example 5d of Chapter 6, in that case,

$$E[Y|X = x] = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$$ ■

7.7 Moment Generating Functions

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

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

We call $M(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $M(t)$ and then evaluating the result at $t = 0$. For example,

$$\begin{aligned} M'(t) &= \frac{d}{dt}E[e^{tX}] \\ &= E\left[\frac{d}{dt}(e^{tX})\right] \\ &= E[Xe^{tX}] \end{aligned} \tag{7.1}$$

where we have assumed that the interchange of the differentiation and expectation operators is legitimate. That is, we have assumed that

$$\frac{d}{dt}\left[\sum_x e^{tx}p(x)\right] = \sum_x \frac{d}{dt}[e^{tx}p(x)]$$

in the discrete case and

$$\frac{d}{dt}\left[\int e^{tx}f(x)\,dx\right] = \int \frac{d}{dt}[e^{tx}f(x)]\,dx$$

in the continuous case. This assumption can almost always be justified and, indeed, is valid for all of the distributions considered in this book. Hence, from Equation (7.1), evaluated at $t = 0$, we obtain

$$M'(0) = E[X]$$

Similarly,

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

Thus,

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

In general, the nth derivative of $M(t)$ is given by

$$M^n(t) = E[X^n e^{tX}] \quad n \geq 1$$

implying that

$$M^n(0) = E[X^n] \quad n \geq 1$$

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

Example 7a **Binomial distribution with parameters n and p**

If X is a binomial random variable with parameters n and p, then

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

where the last equality follows from the binomial theorem. Differentiation yields

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

Thus,

$$E[X] = M'(0) = np$$

Differentiating a second time yields

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

so

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

The variance of X is given by

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

verifying the result obtained previously. ■

Example 7b **Poisson distribution with mean λ**

If X is a Poisson random variable with parameter λ, then

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

Differentiation yields

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

Thus,

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

Hence, both the mean and the variance of the Poisson random variable equal λ. ■

Example 7c **Exponential distribution with parameter λ**

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

We note from this derivation that for the exponential distribution, $M(t)$ is defined only for values of t less than λ. Differentiation of $M(t)$ yields

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

Hence,

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

The variance of X is given by

$$\begin{aligned}\text{Var}(X) &= E[X^2] - (E[X])^2 \\ &= \frac{1}{\lambda^2}\end{aligned}$$

■

Example 7d **Normal distribution**

We first compute the moment generating function of a standard normal random variable with parameters 0 and 1. Letting Z be such a random variable, we have

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

Hence, the moment generating function of the standard normal random variable Z is given by $M_Z(t) = e^{t^2/2}$. To obtain the moment generating function of an arbitrary normal random variable, we recall (see Section 5.4) that $X = \mu + \sigma Z$ will have a normal distribution with parameters μ and σ^2 whenever Z is a standard normal random variable. Hence, the moment generating function of such a random variable is given by

$$\begin{aligned}M_X(t) &= E[e^{tX}] \\ &= E[e^{t(\mu+\sigma Z)}] \\ &= E[e^{t\mu}e^{t\sigma Z}] \\ &= e^{t\mu}E[e^{t\sigma Z}] \\ &= e^{t\mu}M_Z(t\sigma) \\ &= e^{t\mu}e^{(t\sigma)^2/2} \\ &= \exp\left\{\frac{\sigma^2t^2}{2} + \mu t\right\}\end{aligned}$$

By differentiating, we obtain

$$M'_X(t) = (\mu + t\sigma^2)\exp\left\{\frac{\sigma^2t^2}{2} + \mu t\right\}$$

$$M''_X(t) = (\mu + t\sigma^2)^2\exp\left\{\frac{\sigma^2t^2}{2} + \mu t\right\} + \sigma^2\exp\left\{\frac{\sigma^2t^2}{2} + \mu t\right\}$$

Table 7.1 Discrete Probability Distribution.

	Probability mass function, $p(x)$	Moment generating function, $M(t)$	Mean	Variance
Binomial with parameters n, p; $0 \le p \le 1$	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$	$(pe^t + 1 - p)^n$	np	$np(1-p)$
Poisson with parameter $\lambda > 0$	$e^{-\lambda}\frac{\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$	$\exp\{\lambda(e^t - 1)\}$	λ	λ
Geometric with parameter $0 \le p \le 1$	$p(1-p)^{x-1}$ $x = 1, 2, \ldots$	$\frac{pe^t}{1-(1-p)e^t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative binomial with parameters r, p; $0 \le p \le 1$	$\binom{n-1}{r-1} p^r (1-p)^{n-r}$ $n = r, r+1, \ldots$	$\left[\frac{pe^t}{1-(1-p)e^t}\right]^r$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

Thus,

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

implying that

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

■

Tables 7.1 and 7.2 (on page 364) give the moment generating functions for some common discrete and continuous distributions.

An important property of moment generating functions is that the moment generating function of the sum of independent random variables equals the product of the individual moment generating functions. To prove this, suppose that X and Y are independent and have moment generating functions $M_X(t)$ and $M_Y(t)$, respectively. Then $M_{X+Y}(t)$, the moment generating function of $X + Y$, is given by

$$\begin{aligned} M_{X+Y}(t) &= E[e^{t(X+Y)}] \\ &= E[e^{tX} e^{tY}] \\ &= E[e^{tX}]E[e^{tY}] \\ &= M_X(t) M_Y(t) \end{aligned}$$

where the next-to-last equality follows from Proposition 4.1, since X and Y are independent.

Another important result is that the moment generating function uniquely determines the distribution. That is, if $M_X(t)$ exists and is finite in some region about $t = 0$, then the distribution of X is uniquely determined. For instance, if

Table 7.2 Continuous Probability Distribution.

	Probability density function, $f(x)$	Moment generating function, $M(t)$	Mean	Variance
Uniform over (a, b)	$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$\frac{\lambda}{\lambda - t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma with parameters $(s, \lambda), \lambda > 0$	$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{s-1}}{\Gamma(s)} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$\left(\frac{\lambda}{\lambda - t}\right)^s$	$\frac{s}{\lambda}$	$\frac{s}{\lambda^2}$
Normal with parameters (μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$	μ	σ^2

$$M_X(t) = \left(\frac{1}{2}\right)^{10} (e^t + 1)^{10},$$

then it follows from Table 7.1 that X is a binomial random variable with parameters 10 and $\frac{1}{2}$.

Example 7e Suppose that the moment generating function of a random variable X is given by $M(t) = e^{3(e^t-1)}$. What is $P\{X = 0\}$?

Solution We see from Table 7.1 that $M(t) = e^{3(e^t-1)}$ 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 7f **Sums of independent binomial random variables**

If X and Y are independent binomial random variables with parameters (n, p) and (m, p), respectively, what is the distribution of $X + Y$?

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

$$\begin{aligned} M_{X+Y}(t) = M_X(t)M_Y(t) &= (pe^t + 1 - p)^n (pe^t + 1 - p)^m \\ &= (pe^t + 1 - p)^{m+n} \end{aligned}$$

However, $(pe^t + 1 - p)^{m+n}$ is 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 7g **Sums of independent Poisson random variables**

Calculate the distribution of $X + Y$ when X and Y are independent Poisson random variables with means respectively λ_1 and λ_2.

Solution

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= \exp\{\lambda_1(e^t - 1)\} \exp\{\lambda_2(e^t - 1)\} \\ &= \exp\{(\lambda_1 + \lambda_2)(e^t - 1)\} \end{aligned}$$

Hence, $X + Y$ is Poisson distributed with mean $\lambda_1 + \lambda_2$, verifying the result given in Example 3e of Chapter 6. ■

Example 7h **Sums of independent normal random variables**

Show that if X and Y are independent normal random variables with respective parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), then $X + Y$ is normal with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

Solution

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

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

Example 7i Compute the moment generating function of a chi-squared random variable with n degrees of freedom.

Solution We can represent such a random variable as

$$Z_1^2 + \cdots + Z_n^2$$

where $Z_1, \ldots, Z_n$ are independent standard normal random variables. Let $M(t)$ be its moment generating function. Then, by the preceding,

$$M(t) = (E[e^{tZ^2}])^n$$

where Z is a standard normal random variable. Now,

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

where the next-to-last equality uses the fact that the normal density with mean 0 and variance σ^2 integrates to 1. Therefore,

$$M(t) = (1 - 2t)^{-n/2}$$ ■

Example 7j **Moment generating function of the sum of a random number of random variables**

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, and let N be a nonnegative, integer-valued random variable that is independent of the sequence $X_i, i \geq 1$. We want to compute the moment generating function of

$$Y = \sum_{i=1}^{N} X_i$$

(In Example 5d, Y was interpreted as the amount of money spent in a store on a given day when both the amount spent by a customer and the number of customers are random variables.)

To compute the moment generating function of Y, we first condition on N as follows:

$$\begin{aligned} E\left[\exp\left\{t\sum_1^N X_i\right\}\middle| N=n\right] &= E\left[\exp\left\{t\sum_1^n X_i\right\}\middle| N=n\right] \\ &= E\left[\exp\left\{t\sum_1^n X_i\right\}\right] \\ &= [M_X(t)]^n \end{aligned}$$

where

$$M_X(t) = E[e^{tX_i}]$$

Hence,

$$E[e^{tY}|N] = (M_X(t))^N$$

Thus,

$$M_Y(t) = E[(M_X(t))^N]$$

The moments of Y can now be obtained upon differentiation, as follows:

$$M'_Y(t) = E[N(M_X(t))^{N-1}M'_X(t)]$$

So

$$\begin{aligned} E[Y] &= M'_Y(0) \\ &= E[N(M_X(0))^{N-1}M'_X(0)] \\ &= E[NE[X]] \\ &= E[N]E[X] \end{aligned} \tag{7.2}$$

verifying the result of Example 5d. (In this last set of equalities, we have used the fact that $M_X(0) = E[e^{0X}] = 1$.)

Also,

$$M''_Y(t) = E[N(N-1)(M_X(t))^{N-2}(M'_X(t))^2 + N(M_X(t))^{N-1}M''_X(t)]$$

so

$$\begin{aligned} E[Y^2] &= M''_Y(0) \\ &= E[N(N-1)(E[X])^2 + NE[X^2]] \\ &= (E[X])^2(E[N^2] - E[N]) + E[N]E[X^2] \\ &= E[N](E[X^2] - (E[X])^2) + (E[X])^2E[N^2] \\ &= E[N]\text{Var}(X) + (E[X])^2E[N^2] \end{aligned} \tag{7.3}$$

Hence, from Equations (7.2) and (7.3), we have

$$\begin{aligned} \text{Var}(Y) &= E[N]\text{Var}(X) + (E[X])^2(E[N^2] - (E[N])^2) \\ &= E[N]\text{Var}(X) + (E[X])^2\text{Var}(N) \end{aligned}$$

■