Chapter 2

Conditional Probability

2.1 The Definition of Conditional Probability

A major use of probability in statistical inference is the updating of probabilities when certain events are observed. The updated probability of event A after we learn that event B has occurred is the conditional probability of A given B.

Example 2.1.1 Lottery Ticket. Consider a state lottery game in which six numbers are drawn without replacement from a bin containing the numbers 1–30. Each player tries to match the set of six numbers that will be drawn without regard to the order in which the numbers are drawn. Suppose that you hold a ticket in such a lottery with the numbers 1, 14, 15, 20, 23, and 27. You turn on your television to watch the drawing but all you see is one number, 15, being drawn when the power suddenly goes off in your house. You don't even know whether 15 was the first, last, or some in-between draw. However, now that you know that 15 appears in the winning draw, the probability that your ticket is a winner must be higher than it was before you saw the draw. How do you calculate the revised probability? ◀

Example 2.1.1 is typical of the following situation. An experiment is performed for which the sample space S is given (or can be constructed easily) and the probabilities are available for all of the events of interest. We then learn that some event B has occuured, and we want to know how the probability of another event A changes after we learn that B has occurred. In Example 2.1.1, the event that we have learned is $B = \{$one of the numbers drawn is 15$\}$. We are certainly interested in the probability of

$$A = \{\text{the numbers 1, 14, 15, 20, 23, and 27 are drawn}\},$$

and possibly other events.

If we know that the event B has occurred, then we know that the outcome of the experiment is one of those included in B. Hence, to evaluate the probability that A will occur, we must consider the set of those outcomes in B that also result in the occurrence of A. As sketched in Fig. 2.1, this set is precisely the set $A \cap B$. It is therefore natural to calculate the revised probability of A according to the following definition.

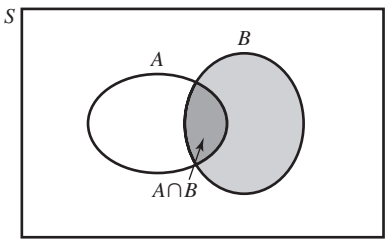

Figure 2.1 The outcomes in the event B that also belong to the event A.

Definition 2.1.1 Conditional Probability. Suppose that we learn that an event B has occurred and that we wish to compute the probability of another event A taking into account that we know that B has occurred. The new probability of A is called the *conditional probability of the event A given that the event B has occurred* and is denoted $\Pr(A|B)$. If $\Pr(B) > 0$, we compute this probability as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{2.1.1}$$

The conditional probability $\Pr(A|B)$ is not defined if $\Pr(B) = 0$.

For convenience, the notation in Definition 2.1.1 is read simply as the conditional probability of A given B. Eq. (2.1.1) indicates that $\Pr(A|B)$ is computed as the proportion of the total probability $\Pr(B)$ that is represented by $\Pr(A \cap B)$, intuitively the proportion of B that is also part of A.

Example 2.1.2 Lottery Ticket. In Example 2.1.1, you learned that the event

$$B = \{\text{one of the numbers drawn is 15}\}$$

has occurred. You want to calculate the probability of the event A that your ticket is a winner. Both events A and B are expressible in the sample space that consists of the $\binom{30}{6} = 30!/(6!24!)$ possible combinations of 30 items taken six at a time, namely, the unordered draws of six numbers from 1–30. The event B consists of combinations that include 15. Since there are 29 remaining numbers from which to choose the other five in the winning draw, there are $\binom{29}{5}$ outcomes in B. It follows that

$$\Pr(B) = \frac{\binom{29}{5}}{\binom{30}{6}} = \frac{29!24!6!}{30!5!24!} = 0.2.$$

The event A that your ticket is a winner consists of a single outcome that is also in B, so $A \cap B = A$, and

$$\Pr(A \cap B) = \Pr(A) = \frac{1}{\binom{30}{6}} = \frac{6!24!}{30!} = 1.68 \times 10^{-6}.$$

It follows that the conditional probability of A given B is

$$\Pr(A|B) = \frac{\frac{6!24!}{30!}}{0.2} = 8.4 \times 10^{-6}.$$

This is five times as large as $\Pr(A)$ before you learned that B had occurred. ◀

Definition 2.1.1 for the conditional probability $\Pr(A|B)$ is worded in terms of the subjective interpretation of probability in Sec. 1.2. Eq. (2.1.1) also has a simple meaning in terms of the frequency interpretation of probability. According to the

frequency interpretation, if an experimental process is repeated a large number of times, then the proportion of repetitions in which the event B will occur is approximately $\Pr(B)$ and the proportion of repetitions in which both the event A and the event B will occur is approximately $\Pr(A \cap B)$. Therefore, among those repetitions in which the event B occurs, the proportion of repetitions in which the event A will also occur is approximately equal to

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Example 2.1.3 Rolling Dice. Suppose that two dice were rolled and it was observed that the sum T of the two numbers was odd. We shall determine the probability that T was less than 8.

If we let A be the event that $T < 8$ and let B be the event that T is odd, then $A \cap B$ is the event that T is 3, 5, or 7. From the probabilities for two dice given at the end of Sec. 1.6, we can evaluate $\Pr(A \cap B)$ and $\Pr(B)$ as follows:

$$\Pr(A \cap B) = \frac{2}{36} + \frac{4}{36} + \frac{6}{36} = \frac{12}{36} = \frac{1}{3},$$

$$\Pr(B) = \frac{2}{36} + \frac{4}{36} + \frac{6}{36} + \frac{4}{36} + \frac{2}{36} = \frac{18}{36} = \frac{1}{2}.$$

Hence,

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{2}{3}.$$ ◀

Example 2.1.4 A Clinical Trial. It is very common for patients with episodes of depression to have a recurrence within two to three years. Prien et al. (1984) studied three treatments for depression: imipramine, lithium carbonate, and a combination. As is traditional in such studies (called *clinical trials*), there was also a group of patients who received a placebo. (A placebo is a treatment that is supposed to be neither helpful nor harmful. Some patients are given a placebo so that they will not know that they did not receive one of the other treatments. None of the other patients knew which treatment or placebo they received, either.) In this example, we shall consider 150 patients who entered the study after an episode of depression that was classified as "unipolar" (meaning that there was no manic disorder). They were divided into the four groups (three treatments plus placebo) and followed to see how many had recurrences of depression. Table 2.1 summarizes the results. If a patient were selected at random from this study and it were found that the patient received the placebo treatment, what is the conditional probability that the patient had a relapse? Let B be the event that the patient received the placebo, and let A be the event that

Table 2.1 Results of the clinical depression study in Example 2.1.4

	Treatment group				
Response	Imipramine	Lithium	Combination	Placebo	*Total*
Relapse	18	13	22	24	77
No relapse	22	25	16	10	73
Total	40	38	38	34	150

the patient had a relapse. We can calculate $\Pr(B) = 34/150$ and $\Pr(A \cap B) = 24/150$ directly from the table. Then $\Pr(A|B) = 24/34 = 0.706$. On the other hand, if the randomly selected patient is found to have received lithium (call this event C) then $\Pr(C) = 38/150$, $\Pr(A \cap C) = 13/150$, and $\Pr(A|C) = 13/38 = 0.342$. Knowing which treatment a patient received seems to make a difference to the probability of relapse. In Chapter 10, we shall study methods for being more precise about how much of a difference it makes. ◀

Example 2.1.5 Rolling Dice Repeatedly. Suppose that two dice are to be rolled repeatedly and the sum T of the two numbers is to be observed for each roll. We shall determine the probability p that the value $T = 7$ will be observed before the value $T = 8$ is observed.

The desired probability p could be calculated directly as follows: We could assume that the sample space S contains all sequences of outcomes that terminate as soon as either the sum $T = 7$ or the sum $T = 8$ is obtained. Then we could find the sum of the probabilities of all the sequences that terminate when the value $T = 7$ is obtained.

However, there is a simpler approach in this example. We can consider the simple experiment in which two dice are rolled. If we repeat the experiment until either the sum $T = 7$ or the sum $T = 8$ is obtained, the effect is to restrict the outcome of the experiment to one of these two values. Hence, the problem can be restated as follows: Given that the outcome of the experiment is either $T = 7$ or $T = 8$, determine the probability p that the outcome is actually $T = 7$.

If we let A be the event that $T = 7$ and let B be the event that the value of T is either 7 or 8, then $A \cap B = A$ and

$$p = \Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)}{\Pr(B)}.$$

From the probabilities for two dice given in Example 1.6.5, $\Pr(A) = 6/36$ and $\Pr(B) = (6/36) + (5/36) = 11/36$. Hence, $p = 6/11$. ◀

The Multiplication Rule for Conditional Probabilities

In some experiments, certain conditional probabilities are relatively easy to assign directly. In these experiments, it is then possible to compute the probability that both of two events occur by applying the next result that follows directly from Eq. (2.1.1) and the analogous definition of $\Pr(B|A)$.

Theorem 2.1.1 Multiplication Rule for Conditional Probabilities. Let A and B be events. If $\Pr(B) > 0$, then

$$\Pr(A \cap B) = \Pr(B) \Pr(A|B).$$

If $\Pr(A) > 0$, then

$$\Pr(A \cap B) = \Pr(A) \Pr(B|A).$$ ■

Example 2.1.6 Selecting Two Balls. Suppose that two balls are to be selected at random, without replacement, from a box containing r red balls and b blue balls. We shall determine the probability p that the first ball will be red and the second ball will be blue.

Let A be the event that the first ball is red, and let B be the event that the second ball is blue. Obviously, $\Pr(A) = r/(r + b)$. Furthermore, if the event A has occurred, then one red ball has been removed from the box on the first draw. Therefore, the

probability of obtaining a blue ball on the second draw will be

$$\Pr(B|A) = \frac{b}{r+b-1}.$$

It follows that

$$\Pr(A \cap B) = \frac{r}{r+b} \cdot \frac{b}{r+b-1}. \quad \blacktriangleleft$$

The principle that has just been applied can be extended to any finite number of events, as stated in the following theorem.

Theorem 2.1.2 Multiplication Rule for Conditional Probabilities. Suppose that $A_1, A_2, \ldots, A_n$ are events such that $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$. Then

$$\begin{aligned} &\Pr(A_1 \cap A_2 \cap \cdots \cap A_n) \\ &\quad = \Pr(A_1) \Pr(A_2|A_1) \Pr(A_3|A_1 \cap A_2) \cdots \Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}). \end{aligned} \tag{2.1.2}$$

Proof The product of probabilities on the right side of Eq. (2.1.2) is equal to

$$\Pr(A_1) \cdot \frac{\Pr(A_1 \cap A_2)}{\Pr(A_1)} \cdot \frac{\Pr(A_1 \cap A_2 \cap A_3)}{\Pr(A_1 \cap A_2)} \cdots \frac{\Pr(A_1 \cap A_2 \cap \cdots \cap A_n)}{\Pr(A_1 \cap A_2 \cdots \cap A_{n-1})}.$$

Since $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$, each of the denominators in this product must be positive. All of the terms in the product cancel each other except the final numerator $\Pr(A_1 \cap A_2 \cap \cdots \cap A_n)$, which is the left side of Eq. (2.1.2). ■

Example 2.1.7 Selecting Four Balls. Suppose that four balls are selected one at a time, without replacement, from a box containing r red balls and b blue balls ($r \geq 2$, $b \geq 2$). We shall determine the probability of obtaining the sequence of outcomes red, blue, red, blue.

If we let R_j denote the event that a red ball is obtained on the jth draw and let B_j denote the event that a blue ball is obtained on the jth draw ($j = 1, \ldots, 4$), then

$$\begin{aligned} \Pr(R_1 \cap B_2 \cap R_3 \cap B_4) &= \Pr(R_1) \Pr(B_2|R_1) \Pr(R_3|R_1 \cap B_2) \Pr(B_4|R_1 \cap B_2 \cap R_3) \\ &= \frac{r}{r+b} \cdot \frac{b}{r+b-1} \cdot \frac{r-1}{r+b-2} \cdot \frac{b-1}{r+b-3}. \quad \blacktriangleleft \end{aligned}$$

Note: Conditional Probabilities Behave Just Like Probabilities. In all of the situations that we shall encounter in this text, every result that we can prove has a conditional version given an event B with $\Pr(B) > 0$. Just replace *all* probabilities by conditional probabilities given B and replace all conditional probabilities given other events C by conditional probabilities given $C \cap B$. For example, Theorem 1.5.3 says that $\Pr(A^c) = 1 - \Pr(A)$. It is easy to prove that $\Pr(A^c|B) = 1 - \Pr(A|B)$ if $\Pr(B) > 0$. (See Exercises 11 and 12 in this section.) Another example is Theorem 2.1.3, which is a conditional version of the multiplication rule Theorem 2.1.2. Although a proof is given for Theorem 2.1.3, we shall not provide proofs of all such conditional theorems, because their proofs are generally very similar to the proofs of the unconditional versions.

Theorem 2.1.3 Suppose that $A_1, A_2, \ldots, A_n, B$ are events such that $\Pr(B) > 0$ and $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}|B) > 0$. Then

$$\begin{aligned} \Pr(A_1 \cap A_2 \cap \cdots \cap A_n|B) = \Pr(A_1|B) \Pr(A_2|A_1 \cap B) \cdots \\ \times \Pr(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1} \cap B). \end{aligned} \tag{2.1.3}$$

Proof The product of probabilities on the right side of Eq. (2.1.3) is equal to

$$\frac{\Pr(A_1 \cap B)}{\Pr(B)} \cdot \frac{\Pr(A_1 \cap A_2 \cap B)}{\Pr(A_1 \cap B)} \cdots \frac{\Pr(A_1 \cap A_2 \cap \cdots \cap A_n B)}{\Pr(A_1 \cap A_2 \cdots \cap A_{n-1} \cap B)}.$$

Since $\Pr(A_1 \cap A_2 \cap \cdots \cap A_{n-1}|B) > 0$, each of the denominators in this product must be positive. All of the terms in the product cancel each other except the first denominator and the final numerator to yield $\Pr(A_1 \cap A_2 \cap \cdots \cap A_n \cap B)/\Pr(B)$, which is the left side of Eq. (2.1.3). ■

Conditional Probability and Partitions

Theorem 1.4.11 shows how to calculate the probability of an event by partitioning the sample space into two events B and B^c. This result easily generalizes to larger partitions, and when combined with Theorem 2.1.1 it leads to a very powerful tool for calculating probabilities.

Definition 2.1.2 Partition. Let S denote the sample space of some experiment, and consider k events $B_1, \ldots, B_k$ in S such that $B_1, \ldots, B_k$ are disjoint and $\bigcup_{i=1}^{k} B_i = S$. It is said that these events form a *partition* of S.

Typically, the events that make up a partition are chosen so that an important source of uncertainty in the problem is reduced if we learn which event has occurred.

Example 2.1.8 Selecting Bolts. Two boxes contain long bolts and short bolts. Suppose that one box contains 60 long bolts and 40 short bolts, and that the other box contains 10 long bolts and 20 short bolts. Suppose also that one box is selected at random and a bolt is then selected at random from that box. We would like to determine the probability that this bolt is long. ◀

Partitions can facilitate the calculations of probabilities of certain events.

Theorem 2.1.4 Law of total probability. Suppose that the events $B_1, \ldots, B_k$ form a partition of the space S and $\Pr(B_j) > 0$ for $j = 1, \ldots, k$. Then, for every event A in S,

$$\Pr(A) = \sum_{j=1}^{k} \Pr(B_j) \Pr(A|B_j). \tag{2.1.4}$$

Proof The events $B_1 \cap A$, $B_2 \cap A, \ldots, B_k \cap A$ will form a partition of A, as illustrated in Fig. 2.2. Hence, we can write

$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_k \cap A).$$

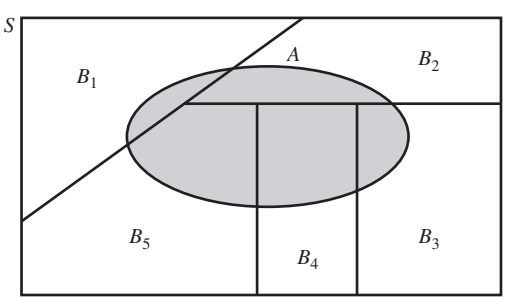

Figure 2.2 The intersections of A with events $B_1, \ldots, B_5$ of a partition in the proof of Theorem 2.1.4.

Furthermore, since the k events on the right side of this equation are disjoint,

$$\Pr(A) = \sum_{j=1}^{k} \Pr(B_j \cap A).$$

Finally, if $\Pr(B_j) > 0$ for $j = 1, \ldots, k$, then $\Pr(B_j \cap A) = \Pr(B_j) \Pr(A|B_j)$ and it follows that Eq. (2.1.4) holds. ■

Example 2.1.9

Selecting Bolts. In Example 2.1.8, let B_1 be the event that the first box (the one with 60 long and 40 short bolts) is selected, let B_2 be the event that the second box (the one with 10 long and 20 short bolts) is selected, and let A be the event that a long bolt is selected. Then

$$\Pr(A) = \Pr(B_1) \Pr(A|B_1) + \Pr(B_2) \Pr(A|B_2).$$

Since a box is selected at random, we know that $\Pr(B_1) = \Pr(B_2) = 1/2$. Furthermore, the probability of selecting a long bolt from the first box is $\Pr(A|B_1) = 60/100 = 3/5$, and the probability of selecting a long bolt from the second box is $\Pr(A|B_2) = 10/30 = 1/3$. Hence,

$$\Pr(A) = \frac{1}{2} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{1}{3} = \frac{7}{15}.$$ ◀

Example 2.1.10

Achieving a High Score. Suppose that a person plays a game in which his score must be one of the 50 numbers $1, 2, \ldots, 50$ and that each of these 50 numbers is equally likely to be his score. The first time he plays the game, his score is X. He then continues to play the game until he obtains another score Y such that $Y \geq X$. We will assume that, conditional on previous plays, the 50 scores remain equally likely on all subsequent plays. We shall determine the probability of the event A that $Y = 50$.

For each $i = 1, \ldots, 50$, let B_i be the event that $X = i$. Conditional on B_i, the value of Y is equally likely to be any one of the numbers $i, i + 1, \ldots, 50$. Since each of these $(51 - i)$ possible values for Y is equally likely, it follows that

$$\Pr(A|B_i) = \Pr(Y = 50|B_i) = \frac{1}{51 - i}.$$

Furthermore, since the probability of each of the 50 values of X is 1/50, it follows that $\Pr(B_i) = 1/50$ for all i and

$$\Pr(A) = \sum_{i=1}^{50} \frac{1}{50} \cdot \frac{1}{51 - i} = \frac{1}{50} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{50}\right) = 0.0900.$$ ◀

Note: Conditional Version of Law of Total Probability. The law of total probability has an analog conditional on another event C, namely,

$$\Pr(A|C) = \sum_{j=1}^{k} \Pr(B_j|C) \Pr(A|B_j \cap C). \tag{2.1.5}$$

The reader can prove this in Exercise 17.

Augmented Experiment In some experiments, it may not be clear from the initial description of the experiment that a partition exists that will facilitate the calculation of probabilities. However, there are many such experiments in which such a partition exists if we imagine that the experiment has some additional structure. Consider the following modification of Examples 2.1.8 and 2.1.9.

Example 2.1.11

Selecting Bolts. There is one box of bolts that contains some long and some short bolts. A manager is unable to open the box at present, so she asks her employees what is the composition of the box. One employee says that it contains 60 long bolts and 40 short bolts. Another says that it contains 10 long bolts and 20 short bolts. Unable to reconcile these opinions, the manager decides that each of the employees is correct with probability 1/2. Let B_1 be the event that the box contains 60 long and 40 short bolts, and let B_2 be the event that the box contains 10 long and 20 short bolts. The probability that the first bolt selected is long is now calculated precisely as in Example 2.1.9. ◀

In Example 2.1.11, there is only one box of bolts, but we believe that it has one of two possible compositions. We let the events B_1 and B_2 determine the possible compositions. This type of situation is very common in experiments.

Example 2.1.12

A Clinical Trial. Consider a clinical trial such as the study of treatments for depression in Example 2.1.4. As in many such trials, each patient has two possible outcomes, in this case relapse and no relapse. We shall refer to relapse as "failure" and no relapse as "success." For now, we shall consider only patients in the imipramine treatment group. If we knew the effectiveness of imipramine, that is, the proportion p of successes among all patients who might receive the treatment, then we might model the patients in our study as having probability p of success. Unfortunately, we do not know p at the start of the trial. In analogy to the box of bolts with unknown composition in Example 2.1.11, we can imagine that the collection of all available patients (from which the 40 imipramine patients in this trial were selected) has two or more possible compositions. We can imagine that the composition of the collection of patients determines the proportion that will be success. For simplicity, in this example, we imagine that there are 11 different possible compositions of the collection of patients. In particular, we assume that the proportions of success for the 11 possible compositions are $0, 1/10, \ldots, 9/10, 1$. (We shall be able to handle more realistic models for p in Chapter 3.) For example, if we knew that our patients were drawn from a collection with the proportion 3/10 of successes, we would be comfortable saying that the patients in our sample each have success probability $p = 3/10$. The value of p is an important source of uncertainty in this problem, and we shall partition the sample space by the possible values of p. For $j = 1, \ldots, 11$, let B_j be the event that our sample was drawn from a collection with proportion $(j-1)/10$ of successes. We can also identify B_j as the event $\{p = (j-1)/10\}$.

Now, let E_1 be the event that the first patient in the imipramine group has a success. We defined each event B_j so that $\Pr(E_1|B_j) = (j-1)/10$. Supppose that, prior to starting the trial, we believe that $\Pr(B_j) = 1/11$ for each j. It follows that

$$\Pr(E_1) = \sum_{j=1}^{11} \frac{1}{11}\frac{j-1}{10} = \frac{55}{110} = \frac{1}{2}, \tag{2.1.6}$$

where the second equality uses the fact that $\sum_{j=1}^{n} j = n(n+1)/2$. ◀

The events $B_1, B_2, \ldots, B_{11}$ in Example 2.1.12 can be thought of in much the same way as the two events B_1 and B_2 that determine the mixture of long and short bolts in Example 2.1.11. There is only one box of bolts, but there is uncertainty about its composition. Similarly in Example 2.1.12, there is only one group of patients, but we believe that it has one of 11 possible compositions determined by the events $B_1, B_2, \ldots, B_{11}$. To call these events, they must be subsets of the sample space for the experiment in question. That will be the case in Example 2.1.12 if we imagine that

the experiment consists not only of observing the numbers of successes and failures among the patients but also of potentially observing enough additional patients to be able to compute p, possibly at some time very far in the future. Similarly, in Example 2.1.11, the two events B_1 and B_2 are subsets of the sample space if we imagine that the experiment consists not only of observing one sample bolt but also of potentially observing the entire composition of the box.

Throughout the remainder of this text, we shall implicitly assume that experiments are augmented to include outcomes that determine the values of quantities such as p. We shall not require that we ever get to observe the complete outcome of the experiment so as to tell us precisely what p is, but merely that there is an experiment that includes all of the events of interest to us, including those that determine quantities like p.

Definition 2.1.3

Augmented Experiment. If desired, any experiment can be augmented to include the potential or hypothetical observation of as much additional information as we would find useful to help us calculate any probabilities that we desire.

Definition 2.1.3 is worded somewhat vaguely because it is intended to cover a wide variety of cases. Here is an explicit application to Example 2.1.12.

Example 2.1.13

A Clinical Trial. In Example 2.1.12, we could explicitly assume that there exists an infinite sequence of patients who could be treated with imipramine even though we will observe only finitely many of them. We could let the sample space consist of infinite sequences of the two symbols S and F such as $(S, S, F, S, F, F, F, \ldots)$. Here S in coordinate i means that the ith patient is a success, and F stands for failure. So, the event E_1 in Example 2.1.12 is the event that the first coordinate is S. The example sequence above is then in the event E_1. To accommodate our interpretation of p as the proportion of successes, we can assume that, for every such sequence, the proportion of S's among the first n coordinates gets close to one of the numbers $0, 1/10, \ldots, 9/10, 1$ as n increases. In this way, p is explicitly the limit of the proportion of successes we would observe if we could find a way to observe indefinitely. In Example 2.1.12, B_2 is the event consisting of all the outcomes in which the limit of the proportion of S's equals $1/10$, B_3 is the set of outcomes in which the limit is $2/10$, etc. Also, we observe only the first 40 coordinates of the infinite sequence, but we still behave as if p exists and could be determined if only we could observe forever. ◀

In the remainder of the text, there will be many experiments that we assume are augmented. In such cases, we will mention which quantities (such as p in Example 2.1.13) would be determined by the augmented part of the experiment even if we do not explicitly mention that the experiment is augmented.

The Game of Craps

We shall conclude this section by discussing a popular gambling game called craps. One version of this game is played as follows: A player rolls two dice, and the sum of the two numbers that appear is observed. If the sum on the first roll is 7 or 11, the player wins the game immediately. If the sum on the first roll is 2, 3, or 12, the player loses the game immediately. If the sum on the first roll is 4, 5, 6, 8, 9, or 10, then the two dice are rolled again and again until the sum is either 7 or the original value. If the original value is obtained a second time before 7 is obtained, then the

player wins. If the sum 7 is obtained before the original value is obtained a second time, then the player loses.

We shall now compute the probability $\Pr(W)$, where W is the event that the player will win. Let the sample space S consist of all possible sequences of sums from the rolls of dice that might occur in a game. For example, some of the elements of S are (4, 7), (11), (4, 3, 4), (12), (10, 8, 2, 12, 6, 7), etc. We see that $(11) \in W$ but $(4, 7) \in W^c$, etc.. We begin by noticing that whether or not an outcome is in W depends in a crucial way on the first roll. For this reason, it makes sense to partition W according to the sum on the first roll. Let B_i be the event that the first roll is i for $i = 2, \ldots, 12$.

Theorem 2.1.4 tells us that $\Pr(W) = \sum_{i=2}^{12} \Pr(B_i) \Pr(W|B_i)$. Since $\Pr(B_i)$ for each i was computed in Example 1.6.5, we need to determine $\Pr(W|B_i)$ for each i. We begin with $i = 2$. Because the player loses if the first roll is 2, we have $\Pr(W|B_2) = 0$. Similarly, $\Pr(W|B_3) = 0 = \Pr(W|B_{12})$. Also, $\Pr(W|B_7) = 1$ because the player wins if the first roll is 7. Similarly, $\Pr(W|B_{11}) = 1$.

For each first roll $i \in \{4, 5, 6, 8, 9, 10\}$, $\Pr(W|B_i)$ is the probability that, in a sequence of dice rolls, the sum i will be obtained before the sum 7 is obtained. As described in Example 2.1.5, this probability is the same as the probability of obtaining the sum i when the sum must be either i or 7. Hence,

$$\Pr(W|B_i) = \frac{\Pr(B_i)}{\Pr(B_i \cup B_7)}.$$

We compute the necessary values here:

$$\Pr(W|B_4) = \frac{\frac{3}{36}}{\frac{3}{36} + \frac{6}{36}} = \frac{1}{3}, \qquad P(W|B_5) = \frac{\frac{4}{36}}{\frac{4}{36} + \frac{6}{36}} = \frac{2}{5},$$

$$\Pr(W|B_6) = \frac{\frac{5}{36}}{\frac{5}{36} + \frac{6}{36}} = \frac{5}{11}, \qquad \Pr(W|B_8) = \frac{\frac{5}{36}}{\frac{5}{36} + \frac{6}{36}} = \frac{5}{11},$$

$$\Pr(W|B_9) = \frac{\frac{4}{36}}{\frac{4}{36} + \frac{6}{36}} = \frac{2}{5}, \qquad \Pr(W|B_{10}) = \frac{\frac{3}{36}}{\frac{3}{36} + \frac{6}{36}} = \frac{1}{3}.$$

Finally, we compute the sum $\sum_{i=2}^{12} \Pr(B_i) \Pr(W|B_i)$:

$$\Pr(W) = \sum_{i=2}^{12} \Pr(B_i) \Pr(W|B_i) = 0 + 0 + \frac{3}{36}\frac{1}{3} + \frac{4}{36}\frac{2}{5} + \frac{5}{36}\frac{5}{11} + \frac{6}{36}$$
$$+ \frac{5}{36}\frac{5}{11} + \frac{4}{36}\frac{2}{5} + \frac{3}{36}\frac{1}{3} + \frac{2}{36} + 0 = \frac{2928}{5940} = 0.493.$$

Thus, the probability of winning in the game of craps is slightly less than 1/2.

Summary

The revised probability of an event A after learning that event B (with $\Pr(B) > 0$) has occurred is the conditional probability of A given B, denoted by $\Pr(A|B)$ and computed as $\Pr(A \cap B)/\Pr(B)$. Often it is easy to assess a conditional probability, such as $\Pr(A|B)$, directly. In such a case, we can use the multiplication rule for conditional probabilities to compute $\Pr(A \cap B) = \Pr(B) \Pr(A|B)$. All probability results have versions conditional on an event B with $\Pr(B) > 0$: Just change *all* probabilities so that they are conditional on B in addition to anything else they were already

conditional on. For example, the multiplication rule for conditional probabilities becomes $\Pr(A_1 \cap A_2|B) = \Pr(A_1|B)\Pr(A_2|A_1 \cap B)$. A partition is a collection of disjoint events whose union is the whole sample space. To be most useful, a partition is chosen so that an important source of uncertainty is reduced if we learn which one of the partition events occurs. If the conditional probability of an event A is available given each event in a partition, the law of total probability tells how to combine these conditional probabilities to get $\Pr(A)$.

Exercises

1. If $A \subset B$ with $\Pr(B) > 0$, what is the value of $\Pr(A|B)$?

2. If A and B are disjoint events and $\Pr(B) > 0$, what is the value of $\Pr(A|B)$?

3. If S is the sample space of an experiment and A is any event in that space, what is the value of $\Pr(A|S)$?

4. Each time a shopper purchases a tube of toothpaste, he chooses either brand A or brand B. Suppose that for each purchase after the first, the probability is 1/3 that he will choose the same brand that he chose on his preceding purchase and the probability is 2/3 that he will switch brands. If he is equally likely to choose either brand A or brand B on his first purchase, what is the probability that both his first and second purchases will be brand A and both his third and fourth purchases will be brand B?

5. A box contains r red balls and b blue balls. One ball is selected at random and its color is observed. The ball is then returned to the box and k additional balls of the same color are also put into the box. A second ball is then selected at random, its color is observed, and it is returned to the box together with k additional balls of the same color. Each time another ball is selected, the process is repeated. If four balls are selected, what is the probability that the first three balls will be red and the fourth ball will be blue?

6. A box contains three cards. One card is red on both sides, one card is green on both sides, and one card is red on one side and green on the other. One card is selected from the box at random, and the color on one side is observed. If this side is green, what is the probability that the other side of the card is also green?

7. Consider again the conditions of Exercise 2 of Sec. 1.10. If a family selected at random from the city subscribes to newspaper A, what is the probability that the family also subscribes to newspaper B?

8. Consider again the conditions of Exercise 2 of Sec. 1.10. If a family selected at random from the city subscribes to at least one of the three newspapers A, B, and C, what is the probability that the family subscribes to newspaper A?

9. Suppose that a box contains one blue card and four red cards, which are labeled A, B, C, and D. Suppose also that two of these five cards are selected at random, without replacement.

a. If it is known that card A has been selected, what is the probability that both cards are red?

b. If it is known that at least one red card has been selected, what is the probability that both cards are red?

10. Consider the following version of the game of craps: The player rolls two dice. If the sum on the first roll is 7 or 11, the player wins the game immediately. If the sum on the first roll is 2, 3, or 12, the player loses the game immediately. However, if the sum on the first roll is 4, 5, 6, 8, 9, or 10, then the two dice are rolled again and again until the sum is either 7 or 11 or the original value. If the original value is obtained a second time before either 7 or 11 is obtained, then the player wins. If either 7 or 11 is obtained before the original value is obtained a second time, then the player loses. Determine the probability that the player will win this game.

11. For any two events A and B with $\Pr(B) > 0$, prove that $\Pr(A^c|B) = 1 - \Pr(A|B)$.

12. For any three events A, B, and D, such that $\Pr(D) > 0$, prove that $\Pr(A \cup B|D) = \Pr(A|D) + \Pr(B|D) - \Pr(A \cap B|D)$.

13. A box contains three coins with a head on each side, four coins with a tail on each side, and two fair coins. If one of these nine coins is selected at random and tossed once, what is the probability that a head will be obtained?

14. A machine produces defective parts with three different probabilities depending on its state of repair. If the machine is in good working order, it produces defective parts with probability 0.02. If it is wearing down, it produces defective parts with probability 0.1. If it needs maintenance, it produces defective parts with probability 0.3. The probability that the machine is in good working order is 0.8, the probability that it is wearing down is 0.1, and the probability that it needs maintenance is 0.1. Compute the probability that a randomly selected part will be defective.

15. The percentages of voters classed as Liberals in three different election districts are divided as follows: in the first district, 21 percent; in the second district, 45 percent; and in the third district, 75 percent. If a district is selected at random and a voter is selected at random from that district, what is the probability that she will be a Liberal?

16. Consider again the shopper described in Exercise 4. On each purchase, the probability that he will choose the same brand of toothpaste that he chose on his preceding purchase is 1/3, and the probability that he will switch brands is 2/3. Suppose that on his first purchase the probability that he will choose brand A is 1/4 and the probability that he will choose brand B is 3/4. What is the probability that his second purchase will be brand B?

17. Prove the conditional version of the law of total probability (2.1.5).

2.2 Independent Events

If learning that B has occurred does not change the probability of A, then we say that A and B are independent. There are many cases in which events A and B are not independent, but they would be independent if we learned that some other event C had occurred. In this case, A and B are conditionally independent given C.

Example 2.2.1 Tossing Coins. Suppose that a fair coin is tossed twice. The experiment has four outcomes, HH, HT, TH, and TT, that tell us how the coin landed on each of the two tosses. We can assume that this sample space is simple so that each outcome has probability 1/4. Suppose that we are interested in the second toss. In particular, we want to calculate the probability of the event $A = \{\text{H on second toss}\}$. We see that $A = \{\text{HH,TH}\}$, so that $\Pr(A) = 2/4 = 1/2$. If we learn that the first coin landed T, we might wish to compute the conditional probability $\Pr(A|B)$ where $B = \{\text{T on first toss}\}$. Using the definition of conditional probability, we easily compute

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/4}{1/2} = \frac{1}{2},$$

because $A \cap B = \{TH\}$ has probability 1/4. We see that $\Pr(A|B) = \Pr(A)$; hence, we don't change the probability of A even after we learn that B has occurred. ◀

Definition of Independence

The conditional probability of the event A given that the event B has occurred is the revised probability of A after we learn that B has occurred. It might be the case, however, that no revision is necessary to the probability of A even after we learn that B occurs. This is precisely what happened in Example 2.2.1. In this case, we say that A and B are *independent events*. As another example, if we toss a coin and then roll a die, we could let A be the event that the die shows 3 and let B be the event that the coin lands with heads up. If the tossing of the coin is done in isolation of the rolling of the die, we might be quite comfortable assigning $\Pr(A|B) = \Pr(A) = 1/6$. In this case, we say that A and B are independent events.

In general, if $\Pr(B) > 0$, the equation $\Pr(A|B) = \Pr(A)$ can be rewritten as $\Pr(A \cap B)/\Pr(B) = \Pr(A)$. If we multiply both sides of this last equation by $\Pr(B)$, we obtain the equation $\Pr(A \cap B) = \Pr(A)\Pr(B)$. In order to avoid the condition $\Pr(B) > 0$, the mathematical definition of the independence of two events is stated as follows:

Definition 2.2.1 Independent Events. Two events A and B are *independent* if

$$\Pr(A \cap B) = \Pr(A)\Pr(B).$$

Suppose that $\Pr(A) > 0$ and $\Pr(B) > 0$. Then it follows easily from the definitions of independence and conditional probability that A and B are independent if and only if $\Pr(A|B) = \Pr(A)$ and $\Pr(B|A) = \Pr(B)$.

Independence of Two Events

If two events A and B are considered to be independent because the events are physically unrelated, and if the probabilities $\Pr(A)$ and $\Pr(B)$ are known, then the definition can be used to assign a value to $\Pr(A \cap B)$.

Example 2.2.2 Machine Operation. Suppose that two machines 1 and 2 in a factory are operated independently of each other. Let A be the event that machine 1 will become inoperative during a given 8-hour period, let B be the event that machine 2 will become inoperative during the same period, and suppose that $\Pr(A) = 1/3$ and $\Pr(B) = 1/4$. We shall determine the probability that at least one of the machines will become inoperative during the given period.

The probability $\Pr(A \cap B)$ that both machines will become inoperative during the period is

$$\Pr(A \cap B) = \Pr(A)\Pr(B) = \left(\frac{1}{3}\right)\left(\frac{1}{4}\right) = \frac{1}{12}.$$

Therefore, the probability $\Pr(A \cup B)$ that at least one of the machines will become inoperative during the period is

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B)\\ &= \frac{1}{3} + \frac{1}{4} - \frac{1}{12} = \frac{1}{2}.\end{aligned}$$

◀

The next example shows that two events A and B, which are physically related, can, nevertheless, satisfy the definition of independence.

Example 2.2.3 Rolling a Die. Suppose that a balanced die is rolled. Let A be the event that an even number is obtained, and let B be the event that one of the numbers 1, 2, 3, or 4 is obtained. We shall show that the events A and B are independent.

In this example, $\Pr(A) = 1/2$ and $\Pr(B) = 2/3$. Furthermore, since $A \cap B$ is the event that either the number 2 or the number 4 is obtained, $\Pr(A \cap B) = 1/3$. Hence, $\Pr(A \cap B) = \Pr(A)\Pr(B)$. It follows that the events A and B are independent events, even though the occurrence of each event depends on the same roll of a die. ◀

The independence of the events A and B in Example 2.2.3 can also be interpreted as follows: Suppose that a person must bet on whether the number obtained on the die will be even or odd, that is, on whether or not the event A will occur. Since three of the possible outcomes of the roll are even and the other three are odd, the person will typically have no preference between betting on an even number and betting on an odd number.

Suppose also that after the die has been rolled, but before the person has learned the outcome and before she has decided whether to bet on an even outcome or on an odd outcome, she is informed that the actual outcome was one of the numbers 1, 2, 3, or 4, i.e., that the event B has occurred. The person now knows that the outcome was 1, 2, 3, or 4. However, since two of these numbers are even and two are odd, the person will typically still have no preference between betting on an even number and betting on an odd number. In other words, the information that the event B has

occurred is of no help to the person who is trying to decide whether or not the event A has occurred.

Independence of Complements In the foregoing discussion of independent events, we stated that if A and B are independent, then the occurrence or nonoccurrence of A should not be related to the occurrence or nonoccurrence of B. Hence, if A and B satisfy the mathematical definition of independent events, then it should also be true that A and B^c are independent events, that A^c and B are independent events, and that A^c and B^c are independent events. One of these results is established in the next theorem.

Theorem 2.2.1 If two events A and B are independent, then the events A and B^c are also independent.

Proof Theorem 1.5.6 says that

$$\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B).$$

Furthermore, since A and B are independent events, $\Pr(A \cap B) = \Pr(A)\Pr(B)$. It now follows that

$$\begin{aligned}\Pr(A \cap B^c) &= \Pr(A) - \Pr(A)\Pr(B) = \Pr(A)[1 - \Pr(B)] \\ &= \Pr(A)\Pr(B^c).\end{aligned}$$

Therefore, the events A and B^c are independent. ■

The proof of the analogous result for the events A^c and B is similar, and the proof for the events A^c and B^c is required in Exercise 2 at the end of this section.

Independence of Several Events

The definition of independent events can be extended to any number of events, $A_1, \ldots, A_k$. Intuitively, if learning that some of these events do or do not occur does not change our probabilities for any events that depend only on the remaining events, we would say that all k events are independent. The mathematical definition is the following analog to Definition 2.2.1.

Definition 2.2.2 (Mutually) Independent Events. The k events $A_1, \ldots, A_k$ are *independent* (or *mutually independent*) if, for every subset $A_{i_1}, \ldots, A_{i_j}$ of j of these events ($j = 2, 3, \ldots, k$),

$$\Pr(A_{i_1} \cap \cdots \cap A_{i_j}) = \Pr(A_{i_1}) \cdots \Pr(A_{i_j}).$$

As an example, in order for three events A, B, and C to be independent, the following four relations must be satisfied:

$$\begin{aligned}\Pr(A \cap B) &= \Pr(A)\Pr(B), \\ \Pr(A \cap C) &= \Pr(A)\Pr(C), \\ \Pr(B \cap C) &= \Pr(B)\Pr(C),\end{aligned} \tag{2.2.1}$$

and

$$\Pr(A \cap B \cap C) = \Pr(A)\Pr(B)\Pr(C). \tag{2.2.2}$$

It is possible that Eq. (2.2.2) will be satisfied, but one or more of the three relations (2.2.1) will not be satisfied. On the other hand, as is shown in the next example,

it is also possible that each of the three relations (2.2.1) will be satisfied but Eq. (2.2.2) will not be satisfied.

Example 2.2.4 Pairwise Independence. Suppose that a fair coin is tossed twice so that the sample space $S = \{\text{HH, HT, TH, TT}\}$ is simple. Define the following three events:

$$A = \{\text{H on first toss}\} = \{\text{HH, HT}\},$$
$$B = \{\text{H on second toss}\} = \{\text{HH, TH}\}, \text{ and}$$
$$C = \{\text{Both tosses the same}\} = \{\text{HH, TT}\}.$$

Then $A \cap B = A \cap C = B \cap C = A \cap B \cap C = \{HH\}$. Hence,

$$\Pr(A) = \Pr(B) = \Pr(C) = 1/2$$

and

$$\Pr(A \cap B) = \Pr(A \cap C) = \Pr(B \cap C) = \Pr(A \cap B \cap C) = 1/4.$$

It follows that each of the three relations of Eq. (2.2.1) is satisfied but Eq. (2.2.2) is not satisfied. These results can be summarized by saying that the events A, B, and C are *pairwise independent*, but all three events are not independent. ◀

We shall now present some examples that will illustrate the power and scope of the concept of independence in the solution of probability problems.

Example 2.2.5 Inspecting Items. Suppose that a machine produces a defective item with probability p $(0 < p < 1)$ and produces a nondefective item with probability $1 - p$. Suppose further that six items produced by the machine are selected at random and inspected, and that the results (defective or nondefective) for these six items are independent. We shall determine the probability that exactly two of the six items are defective.

It can be assumed that the sample space S contains all possible arrangements of six items, each one of which might be either defective or nondefective. For $j = 1, \ldots, 6$, we shall let D_j denote the event that the jth item in the sample is defective so that D_j^c is the event that this item is nondefective. Since the outcomes for the six different items are independent, the probability of obtaining any particular sequence of defective and nondefective items will simply be the product of the individual probabilities for the items. For example,

$$\begin{aligned}\Pr(D_1^c \cap D_2 \cap D_3^c \cap D_4^c \cap D_5 \cap D_6^c) &= \Pr(D_1^c)\Pr(D_2)\Pr(D_3^c)\Pr(D_4^c)\Pr(D_5)\Pr(D_6^c) \\ &= (1-p)p(1-p)(1-p)p(1-p) = p^2(1-p)^4.\end{aligned}$$

It can be seen that the probability of any other particular sequence in S containing two defective items and four nondefective items will also be $p^2(1-p)^4$. Hence, the probability that there will be exactly two defectives in the sample of six items can be found by multiplying the probability $p^2(1-p)^4$ of any particular sequence containing two defectives by the possible number of such sequences. Since there are $\binom{6}{2}$ distinct arrangements of two defective items and four nondefective items, the probability of obtaining exactly two defectives is $\binom{6}{2}p^2(1-p)^4$. ◀

Example 2.2.6 Obtaining a Defective Item. For the conditions of Example 2.2.5, we shall now determine the probability that at least one of the six items in the sample will be defective.

Since the outcomes for the different items are independent, the probability that all six items will be nondefective is $(1-p)^6$. Therefore, the probability that at least one item will be defective is $1 - (1-p)^6$. ◀

Example 2.2.7 **Tossing a Coin Until a Head Appears.** Suppose that a fair coin is tossed until a head appears for the first time, and assume that the outcomes of the tosses are independent. We shall determine the probability p_n that exactly n tosses will be required.

The desired probability is equal to the probability of obtaining $n-1$ tails in succession and then obtaining a head on the next toss. Since the outcomes of the tosses are independent, the probability of this particular sequence of n outcomes is $p_n = (1/2)^n$.

The probability that a head will be obtained sooner or later (or, equivalently, that tails will not be obtained forever) is

$$\sum_{n=1}^{\infty} p_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1.$$

Since the sum of the probabilities p_n is 1, it follows that the probability of obtaining an infinite sequence of tails without ever obtaining a head must be 0. ◀

Example 2.2.8 **Inspecting Items One at a Time.** Consider again a machine that produces a defective item with probability p and produces a nondefective item with probability $1-p$. Suppose that items produced by the machine are selected at random and inspected one at a time until exactly five defective items have been obtained. We shall determine the probability p_n that exactly n items ($n \geq 5$) must be selected to obtain the five defectives.

The fifth defective item will be the nth item that is inspected if and only if there are exactly four defectives among the first $n-1$ items and then the nth item is defective. By reasoning similar to that given in Example 2.2.5, it can be shown that the probability of obtaining exactly four defectives and $n-5$ nondefectives among the first $n-1$ items is $\binom{n-1}{4} p^4 (1-p)^{n-5}$. The probability that the nth item will be defective is p. Since the first event refers to outcomes for only the first $n-1$ items and the second event refers to the outcome for only the nth item, these two events are independent. Therefore, the probability that both events will occur is equal to the product of their probabilities. It follows that

$$p_n = \binom{n-1}{4} p^5 (1-p)^{n-5}. \quad \blacktriangleleft$$

Example 2.2.9 **People v. Collins.** Finkelstein and Levin (1990) describe a criminal case whose verdict was overturned by the Supreme Court of California in part due to a probability calculation involving both conditional probability and independence. The case, *People v. Collins*, 68 Cal. 2d 319, 438 P.2d 33 (1968), involved a purse snatching in which witnesses claimed to see a young woman with blond hair in a ponytail fleeing from the scene in a yellow car driven by a black man with a beard. A couple meeting the description was arrested a few days after the crime, but no physical evidence was found. A mathematician calculated the probability that a randomly selected couple would possess the described characteristics as about 8.3×10^{-8}, or 1 in 12 million. Faced with such overwhelming odds and no physical evidence, the jury decided that the defendants must have been the only such couple and convicted them. The Supreme Court thought that a more useful probability should have been calculated. Based on the testimony of the witnesses, there was a couple that met the above description. Given that there was already one couple who met the description, what is the conditional probability that there was also a second couple such as the defendants?

Let p be the probability that a randomly selected couple from a population of n couples has certain characteristics. Let A be the event that at least one couple in the population has the characteristics, and let B be the event that at least two couples

have the characteristics. What we seek is $\Pr(B|A)$. Since $B \subset A$, it follows that

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(B)}{\Pr(A)}.$$

We shall calculate $\Pr(B)$ and $\Pr(A)$ by breaking each event into more manageable pieces. Suppose that we number the n couples in the population from 1 to n. Let A_i be the event that couple number i has the characteristics in question for $i = 1, \ldots, n$, and let C be the event that exactly one couple has the characteristics. Then

$$A = (A_1^c \cap A_2^c \cdots \cap A_n^c)^c,$$
$$C = (A_1 \cap A_2^c \cdots \cap A_n^c) \cup (A_1^c \cap A_2 \cap A_3^c \cdots \cap A_n^c) \cup \cdots \cup (A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n),$$
$$B = A \cap C^c.$$

Assuming that the n couples are mutually independent, $\Pr(A^c) = (1-p)^n$, and $\Pr(A) = 1 - (1-p)^n$. The n events whose union is C are disjoint and each one has probability $p(1-p)^{n-1}$, so $\Pr(C) = np(1-p)^{n-1}$. Since $A = B \cup C$ with B and C disjoint, we have

$$\Pr(B) = \Pr(A) - \Pr(C) = 1 - (1-p)^n - np(1-p)^{n-1}.$$

So,

$$\Pr(B|A) = \frac{1 - (1-p)^n - np(1-p)^{n-1}}{1 - (1-p)^n}. \tag{2.2.3}$$

The Supreme Court of California reasoned that, since the crime occurred in a heavily populated area, n would be in the millions. For example, with $p = 8.3 \times 10^{-8}$ and $n = 8{,}000{,}000$, the value of (2.2.3) is 0.2966. Such a probability suggests that there is a reasonable chance that there was another couple meeting the same description as the witnesses provided. Of course, the court did not know how large n was, but the fact that (2.2.3) could easily be so large was grounds enough to rule that reasonable doubt remained as to the guilt of the defendants. ◀

Independence and Conditional Probability Two events A and B with positive probability are independent if and only if $\Pr(A|B) = \Pr(A)$. Similar results hold for larger collections of independent events. The following theorem, for example, is straightforward to prove based on the definition of independence.

Theorem 2.2.2 Let $A_1, \ldots, A_k$ be events such that $\Pr(A_1 \cap \cdots \cap A_k) > 0$. Then $A_1, \ldots, A_k$ are independent if and only if, for every two disjoint subsets $\{i_1, \ldots, i_m\}$ and $\{j_1, \ldots, j_\ell\}$ of $\{1, \ldots, k\}$, we have

$$\Pr(A_{i_1} \cap \cdots \cap A_{i_m} | A_{j_1} \cap \cdots \cap A_{j_\ell}) = \Pr(A_{i_1} \cap \cdots \cap A_{i_m}).$$ ■

Theorem 2.2.2 says that k events are independent if and only if learning that some of the events occur does not change the probability that any combination of the other events occurs.

The Meaning of Independence We have given a mathematical definition of independent events in Definition 2.2.1. We have also given some interpretations for what it means for events to be independent. The most instructive interpretation is the one based on conditional probability. If learning that B occurs does not change the probability of A, then A and B are independent. In simple examples such as tossing what we believe to be a fair coin, we would generally not expect to change our minds

about what is likely to happen on later flips after we observe earlier flips; hence, we declare the events that concern different flips to be independent. However, consider a situation similar to Example 2.2.5 in which items produced by a machine are inspected to see whether or not they are defective. In Example 2.2.5, we declared that the different items were independent and that each item had probability p of being defective. This might make sense if we were confident that we knew how well the machine was performing. But if we were unsure of how the machine were performing, we could easily imagine changing our mind about the probability that the 10th item is defective depending on how many of the first nine items are defective. To be specific, suppose that we begin by thinking that the probability is 0.08 that an item will be defective. If we observe one or zero defective items in the first nine, we might not make much revision to the probability that the 10th item is defective. On the other hand, if we observe eight or nine defectives in the first nine items, we might be uncomfortable keeping the probability at 0.08 that the 10th item will be defective. In summary, when deciding whether to model events as independent, try to answer the following question: "If I were to learn that some of these events occurred, would I change the probabilities of any of the others?" If we feel that we already know everything that we could learn from these events about how likely the others should be, we can safely model them as independent. If, on the other hand, we feel that learning some of these events could change our minds about how likely some of the others are, then we should be more careful about determining the conditional probabilities and not model the events as independent.

Mutually Exclusive Events and Mutually Independent Events Two similar-sounding definitions have appeared earlier in this text. Definition 1.4.10 defines mutually exclusive events, and Definition 2.2.2 defines mutually independent events. It is almost never the case that the same set of events satisfies both definitions. The reason is that if events are disjoint (mutually exclusive), then learning that one occurs means that the others definitely did not occur. Hence, learning that one occurs would change the probabilities for all the others to 0, unless the others already had probability 0. Indeed, this suggests the only condition in which the two definitions would both apply to the same collection of events. The proof of the following result is left to Exercise 24 in this section.

Theorem 2.2.3 Let $n > 1$ and let $A_1, \ldots, A_n$ be events that are mutually exclusive. The events are also mutually independent if and only if all the events except possibly one of them has probability 0. ■

Conditionally Independent Events

Conditional probability and independence combine into one of the most versatile models of data collection. The idea is that, in many circumstances, we are unwilling to say that certain events are independent because we believe that learning some of them will provide information about how likely the others are to occur. But if we knew the frequency with which such events would occur, we might then be willing to assume that they are independent. This model can be illustrated using one of the examples from earlier in this section.

Example 2.2.10 Inspecting Items. Consider again the situation in Example 2.2.5. This time, however, suppose that we believe that we would change our minds about the probabilities of later items being defective were we to learn that certain numbers of early items

were defective. Suppose that we think of the number p from Example 2.2.5 as the proportion of defective items that we would expect to see if we were to inspect a very large sample of items. If we knew this proportion p, and if we were to sample only a few, say, six or 10 items now, we might feel confident maintaining that the probability of a later item being defective remains p even after we inspect some of the earlier items. On the other hand, if we are not sure what would be the proportion of defective items in a large sample, we might not feel confident keeping the probability the same as we continue to inspect.

To be precise, suppose that we treat the proportion p of defective items as unknown and that we are dealing with an augmented experiment as described in Definition 2.1.3. For simplicity, suppose that p can take one of two values, either 0.01 or 0.4, the first corresponding to normal operation and the second corresponding to a need for maintenance. Let B_1 be the event that $p = 0.01$, and let B_2 be the event that $p = 0.4$. If we knew that B_1 had occurred, then we would proceed under the assumption that the events $D_1, D_2, \ldots$ were independent with $\Pr(D_i|B_1) = 0.01$ for all i. For example, we could do the same calculations as in Examples 2.2.5 and 2.2.8 with $p = 0.01$. Let A be the event that we observe exactly two defectives in a random sample of six items. Then $\Pr(A|B_1) = \binom{6}{2}0.01^2 0.99^4 = 1.44 \times 10^{-3}$. Similarly, if we knew that B_2 had occurred, then we would assume that $D_1, D_2, \ldots$ were independent with $\Pr(D_i|B_2) = 0.4$. In this case, $\Pr(A|B_2) = \binom{6}{2}0.4^2 0.6^4 = 0.311$. ◀

In Example 2.2.10, there is no reason that p must be required to assume at most two different values. We could easily allow p to take a third value or a fourth value, etc. Indeed, in Chapter 3 we shall learn how to handle the case in which every number between 0 and 1 is a possible value of p. The point of the simple example is to illustrate the concept of assuming that events are independent conditional on another event, such as B_1 or B_2 in the example.

The formal concept illustrated in Example 2.2.10 is the following:

Definition 2.2.3 Conditional Independence. We say that events $A_1, \ldots, A_k$ are *conditionally independent given* B if, for every subcollection $A_{i_1}, \ldots, A_{i_j}$ of j of these events ($j = 2, 3, \ldots, k$),

$$\Pr\left(A_{i_1} \cap \cdots \cap A_{i_j} \middle| B\right) = \Pr(A_{i_1}|B) \cdots \Pr(A_{i_j}|B).$$

Definition 2.2.3 is identical to Definition 2.2.2 for independent events with the modification that *all* probabilities in the definition are now conditional on B. As a note, even if we assume that events $A_1, \ldots, A_k$ are conditionally independent given B, it is *not* necessary that they be conditionally independent given B^c. In Example 2.2.10, the events $D_1, D_2, \ldots$ were conditionally independent given both B_1 and $B_2 = B_1^c$, which is the typical situation. Exercise 16 in Sec. 2.3 is an example in which events are conditionally independent given one event B but are not conditionally independent given the complement B^c.

Recall that two events A_1 and A_2 (with $\Pr(A_1) > 0$) are independent if and only if $\Pr(A_2|A_1) = \Pr(A_2)$. A similar result holds for conditionally independent events.

Theorem 2.2.4 Suppose that A_1, A_2, and B are events such that $\Pr(A_1 \cap B) > 0$. Then A_1 and A_2 are conditionally independent given B if and only if $\Pr(A_2|A_1 \cap B) = \Pr(A_2|B)$. ■

This is another example of the claim we made earlier that every result we can prove has an analog conditional on an event B. The reader can prove this theorem in Exercise 22.

The Collector's Problem

Suppose that n balls are thrown in a random manner into r boxes ($r \le n$). We shall assume that the n throws are independent and that each of the r boxes is equally likely to receive any given ball. The problem is to determine the probability p that every box will receive at least one ball. This problem can be reformulated in terms of a collector's problem as follows: Suppose that each package of bubble gum contains the picture of a baseball player, that the pictures of r different players are used, that the picture of each player is equally likely to be placed in any given package of gum, and that pictures are placed in different packages independently of each other. The problem now is to determine the probability p that a person who buys n packages of gum ($n \ge r$) will obtain a complete set of r different pictures.

For $i = 1, \ldots, r$, let A_i denote the event that the picture of player i is missing from all n packages. Then $\bigcup_{i=1}^{r} A_i$ is the event that the picture of at least one player is missing. We shall find $\Pr(\bigcup_{i=1}^{r} A_i)$ by applying Eq. (1.10.6).

Since the picture of each of the r players is equally likely to be placed in any particular package, the probability that the picture of player i will not be obtained in any particular package is $(r-1)/r$. Since the packages are filled independently, the probability that the picture of player i will not be obtained in any of the n packages is $[(r-1)/r]^n$. Hence,

$$\Pr(A_i) = \left(\frac{r-1}{r}\right)^n \quad \text{for } i = 1, \ldots, r.$$

Now consider any two players i and j. The probability that neither the picture of player i nor the picture of player j will be obtained in any particular package is $(r-2)/r$. Therefore, the probability that neither picture will be obtained in any of the n packages is $[(r-2)/r]^n$. Thus,

$$\Pr(A_i \cap A_j) = \left(\frac{r-2}{r}\right)^n.$$

If we next consider any three players i, j, and k, we find that

$$\Pr(A_i \cap A_j \cap A_k) = \left(\frac{r-3}{r}\right)^n.$$

By continuing in this way, we finally arrive at the probability $\Pr(A_1 \cap A_2 \cap \cdots \cap A_r)$ that the pictures of all r players are missing from the n packages. Of course, this probability is 0. Therefore, by Eq. (1.10.6) of Sec. 1.10,

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{r} A_i\right) &= r\left(\frac{r-1}{r}\right)^n - \binom{r}{2}\left(\frac{r-2}{r}\right)^n + \cdots + (-1)^r \binom{r}{r-1}\left(\frac{1}{r}\right)^n \\ &= \sum_{j=1}^{r-1} (-1)^{j+1} \binom{r}{j}\left(1 - \frac{j}{r}\right)^n.\end{aligned}$$

Since the probability p of obtaining a complete set of r different pictures is equal to $1 - \Pr(\bigcup_{i=1}^{r} A_i)$, it follows from the foregoing derivation that p can be written in the form

$$p = \sum_{j=0}^{r-1} (-1)^j \binom{r}{j}\left(1 - \frac{j}{r}\right)^n.$$

Summary

A collection of events is independent if and only if learning that some of them occur does not change the probabilities that any combination of the rest of them occurs. Equivalently, a collection of events is independent if and only if the probability of the intersection of every subcollection is the product of the individual probabilities. The concept of independence has a version conditional on another event. A collection of events is independent conditional on B if and only if the conditional probability of the intersection of every subcollection given B is the product of the individual conditional probabilities given B. Equivalently, a collection of events is conditionally independent given B if and only if learning that some of them (and B) occur does not change the conditional probabilities given B that any combination of the rest of them occur. The full power of conditional independence will become more apparent after we introduce Bayes' theorem in the next section.

Exercises

1. If A and B are independent events and $\Pr(B) < 1$, what is the value of $\Pr(A^c|B^c)$?

2. Assuming that A and B are independent events, prove that the events A^c and B^c are also independent.

3. Suppose that A is an event such that $\Pr(A) = 0$ and that B is any other event. Prove that A and B are independent events.

4. Suppose that a person rolls two balanced dice three times in succession. Determine the probability that on each of the three rolls, the sum of the two numbers that appear will be 7.

5. Suppose that the probability that the control system used in a spaceship will malfunction on a given flight is 0.001. Suppose further that a duplicate, but completely independent, control system is also installed in the spaceship to take control in case the first system malfunctions. Determine the probability that the spaceship will be under the control of either the original system or the duplicate system on a given flight.

6. Suppose that 10,000 tickets are sold in one lottery and 5000 tickets are sold in another lottery. If a person owns 100 tickets in each lottery, what is the probability that she will win at least one first prize?

7. Two students A and B are both registered for a certain course. Assume that student A attends class 80 percent of the time, student B attends class 60 percent of the time, and the absences of the two students are independent.

- **a.** What is the probability that at least one of the two students will be in class on a given day?
- **b.** If at least one of the two students is in class on a given day, what is the probability that A is in class that day?

8. If three balanced dice are rolled, what is the probability that all three numbers will be the same?

9. Consider an experiment in which a fair coin is tossed until a head is obtained for the first time. If this experiment is performed three times, what is the probability that exactly the same number of tosses will be required for each of the three performances?

10. The probability that any child in a certain family will have blue eyes is 1/4, and this feature is inherited independently by different children in the family. If there are five children in the family and it is known that at least one of these children has blue eyes, what is the probability that at least three of the children have blue eyes?

11. Consider the family with five children described in Exercise 10.

- **a.** If it is known that the youngest child in the family has blue eyes, what is the probability that at least three of the children have blue eyes?
- **b.** Explain why the answer in part (a) is different from the answer in Exercise 10.

12. Suppose that A, B, and C are three independent events such that $\Pr(A) = 1/4$, $\Pr(B) = 1/3$, and $\Pr(C) = 1/2$. **(a)** Determine the probability that none of these three events will occur. **(b)** Determine the probability that exactly one of these three events will occur.

13. Suppose that the probability that any particle emitted by a radioactive material will penetrate a certain shield is 0.01. If 10 particles are emitted, what is the probability that exactly one of the particles will penetrate the shield?

14. Consider again the conditions of Exercise 13. If 10 particles are emitted, what is the probability that at least one of the particles will penetrate the shield?

15. Consider again the conditions of Exercise 13. How many particles must be emitted in order for the probability to be at least 0.8 that at least one particle will penetrate the shield?

16. In the World Series of baseball, two teams A and B play a sequence of games against each other, and the first team that wins a total of four games becomes the winner of the World Series. If the probability that team A will win any particular game against team B is 1/3, what is the probability that team A will win the World Series?

17. Two boys A and B throw a ball at a target. Suppose that the probability that boy A will hit the target on any throw is 1/3 and the probability that boy B will hit the target on any throw is 1/4. Suppose also that boy A throws first and the two boys take turns throwing. Determine the probability that the target will be hit for the first time on the third throw of boy A.

18. For the conditions of Exercise 17, determine the probability that boy A will hit the target before boy B does.

19. A box contains 20 red balls, 30 white balls, and 50 blue balls. Suppose that 10 balls are selected at random one at a time, with replacement; that is, each selected ball is replaced in the box before the next selection is made. Determine the probability that at least one color will be missing from the 10 selected balls.

20. Suppose that $A_1, \ldots, A_k$ form a sequence of k independent events. Let $B_1, \ldots, B_k$ be another sequence of k events such that for each value of j ($j = 1, \ldots, k$), either $B_j = A_j$ or $B_j = A_j^c$. Prove that $B_1, \ldots, B_k$ are also independent events. *Hint:* Use an induction argument based on the number of events B_j for which $B_j = A_j^c$.

21. Prove Theorem 2.2.2 on page 71. *Hint:* The "only if" direction is direct from the definition of independence on page 68. For the "if" direction, use induction on the value of j in the definition of independence. Let $m = j - 1$ and let $\ell = 1$ with $j_1 = i_j$.

22. Prove Theorem 2.2.4 on page 73.

23. A programmer is about to attempt to compile a series of 11 similar programs. Let A_i be the event that the ith program compiles successfully for $i = 1, \ldots, 11$. When the programming task is easy, the programmer expects that 80 percent of programs should compile. When the programming task is difficult, she expects that only 40 percent of the programs will compile. Let B be the event that the programming task was easy. The programmer believes that the events $A_1, \ldots, A_{11}$ are conditionally independent given B and given B^c.

a. Compute the probability that exactly 8 out of 11 programs will compile given B.

b. Compute the probability that exactly 8 out of 11 programs will compile given B^c.

24. Prove Theorem 2.2.3 on page 72.

2.3 Bayes' Theorem

Suppose that we are interested in which of several disjoint events $B_1, \ldots, B_k$ will occur and that we will get to observe some other event A. If $\Pr(A|B_i)$ is available for each i, then Bayes' theorem is a useful formula for computing the conditional probabilities of the B_i events given A.

We begin with a typical example.

Example 2.3.1

Test for a Disease. Suppose that you are walking down the street and notice that the Department of Public Health is giving a free medical test for a certain disease. The test is 90 percent reliable in the following sense: If a person has the disease, there is a probability of 0.9 that the test will give a positive response; whereas, if a person does not have the disease, there is a probability of only 0.1 that the test will give a positive response.

Data indicate that your chances of having the disease are only 1 in 10,000. However, since the test costs you nothing, and is fast and harmless, you decide to stop and take the test. A few days later you learn that you had a positive response to the test. Now, what is the probability that you have the disease? ◀

The last question in Example 2.3.1 is a prototype of the question for which Bayes' theorem was designed. We have at least two disjoint events ("you have the disease" and "you do not have the disease") about which we are uncertain, and we learn a piece of information (the result of the test) that tells us something about the uncertain events. Then we need to know how to revise the probabilities of the events in the light of the information we learned.

We now present the general structure in which Bayes' theorem operates before returning to the example.

Statement, Proof, and Examples of Bayes' Theorem

Example 2.3.2

Selecting Bolts. Consider again the situation in Example 2.1.8, in which a bolt is selected at random from one of two boxes. Suppose that we cannot tell without making a further effort from which of the two boxes the one bolt is being selected. For example, the boxes may be identical in appearance or somebody else may actually select the box, but we only get to see the bolt. Prior to selecting the bolt, it was equally likely that each of the two boxes would be selected. However, if we learn that event A has occurred, that is, a long bolt was selected, we can compute the conditional probabilities of the two boxes given A. To remind the reader, B_1 is the event that the box is selected containing 60 long bolts and 40 short bolts, while B_2 is the event that the box is selected containing 10 long bolts and 20 short bolts. In Example 2.1.9, we computed $\Pr(A) = 7/15$, $\Pr(A|B_1) = 3/5$, $\Pr(A|B_2) = 1/3$, and $\Pr(B_1) = \Pr(B_2) = 1/2$. So, for example,

$$\Pr(B_1|A) = \frac{\Pr(A \cap B_1)}{\Pr(A)} = \frac{\Pr(B_1)\Pr(A|B_1)}{\Pr(A)} = \frac{\frac{1}{2} \times \frac{3}{5}}{\frac{7}{15}} = \frac{9}{14}.$$

Since the first box has a higher proportion of long bolts than the second box, it seems reasonable that the probability of B_1 should rise after we learn that a long bolt was selected. It must be that $\Pr(B_2|A) = 5/14$ since one or the other box had to be selected. ◀

In Example 2.3.2, we started with uncertainty about which of two boxes would be chosen and then we observed a long bolt drawn from the chosen box. Because the two boxes have different chances of having a long bolt drawn, the observation of a long bolt changed the probabilities of each of the two boxes having been chosen. The precise calculation of how the probabilities change is the purpose of Bayes' theorem.

Theorem 2.3.1

Bayes' theorem. Let the events $B_1, \ldots, B_k$ form a partition of the space S such that $\Pr(B_j) > 0$ for $j = 1, \ldots, k$, and let A be an event such that $\Pr(A) > 0$. Then, for $i = 1, \ldots, k$,

$$\Pr(B_i|A) = \frac{\Pr(B_i)\Pr(A|B_i)}{\sum_{j=1}^{k} \Pr(B_j)\Pr(A|B_j)}. \tag{2.3.1}$$

Proof By the definition of conditional probability,

$$\Pr(B_i|A) = \frac{\Pr(B_i \cap A)}{\Pr(A)}.$$

The numerator on the right side of Eq. (2.3.1) is equal to $\Pr(B_i \cap A)$ by Theorem 2.1.1. The denominator is equal to $\Pr(A)$ according to Theorem 2.1.4. ■

Example 2.3.3

Test for a Disease. Let us return to the example with which we began this section. We have just received word that we have tested positive for a disease. The test was 90 percent reliable in the sense that we described in Example 2.3.1. We want to know the probability that we have the disease after we learn that the result of the test is positive. Some readers may feel that this probability should be about 0.9. However, this feeling completely ignores the small probability of 0.0001 that you had the disease before taking the test. We shall let B_1 denote the event that you have the disease, and let B_2 denote the event that you do not have the disease. The events B_1 and B_2 form a partition. Also, let A denote the event that the response to the test is positive. The event A is information we will learn that tells us something about the partition elements. Then, by Bayes' theorem,

$$\Pr(B_1|A) = \frac{\Pr(A|B_1)\Pr(B_1)}{\Pr(A|B_1)\Pr(B_1) + \Pr(A|B_2)\Pr(B_2)}$$
$$= \frac{(0.9)(0.0001)}{(0.9)(0.0001) + (0.1)(0.9999)} = 0.00090.$$

Thus, the conditional probability that you have the disease given the test result is approximately only 1 in 1000. Of course, this conditional probability is approximately 9 times as great as the probability was before you were tested, but even the conditional probability is quite small.

Another way to explain this result is as follows: Only one person in every 10,000 actually has the disease, but the test gives a positive response for approximately one person in every 10. Hence, the number of positive responses is approximately 1000 times the number of persons who actually have the disease. In other words, out of every 1000 persons for whom the test gives a positive response, only one person actually has the disease. This example illustrates not only the use of Bayes' theorem but also the importance of taking into account all of the information available in a problem. ◀

Example 2.3.4

Identifying the Source of a Defective Item. Three different machines M_1, M_2, and M_3 were used for producing a large batch of similar manufactured items. Suppose that 20 percent of the items were produced by machine M_1, 30 percent by machine M_2, and 50 percent by machine M_3. Suppose further that 1 percent of the items produced by machine M_1 are defective, that 2 percent of the items produced by machine M_2 are defective, and that 3 percent of the items produced by machine M_3 are defective. Finally, suppose that one item is selected at random from the entire batch and it is found to be defective. We shall determine the probability that this item was produced by machine M_2.

Let B_i be the event that the selected item was produced by machine M_i ($i = 1, 2, 3$), and let A be the event that the selected item is defective. We must evaluate the conditional probability $\Pr(B_2|A)$.

The probability $\Pr(B_i)$ that an item selected at random from the entire batch was produced by machine M_i is as follows, for $i = 1, 2, 3$:

$$\Pr(B_1) = 0.2, \quad \Pr(B_2) = 0.3, \quad \Pr(B_3) = 0.5.$$

Furthermore, the probability $\Pr(A|B_i)$ that an item produced by machine M_i will be defective is

$$\Pr(A|B_1) = 0.01, \quad \Pr(A|B_2) = 0.02, \quad \Pr(A|B_3) = 0.03.$$

It now follows from Bayes' theorem that

$$\Pr(B_2|A) = \frac{\Pr(B_2)\Pr(A|B_2)}{\sum_{j=1}^{3}\Pr(B_j)\Pr(A|B_j)}$$
$$= \frac{(0.3)(0.02)}{(0.2)(0.01)+(0.3)(0.02)+(0.5)(0.03)} = 0.26. \quad \blacktriangleleft$$

Example 2.3.5

Identifying Genotypes. Consider a gene that has two alleles (see Example 1.6.4 on page 23) A and a. Suppose that the gene exhibits itself through a trait (such as hair color or blood type) with two versions. We call A *dominant* and a *recessive* if individuals with genotypes AA and Aa have the same version of the trait and the individuals with genotype aa have the other version. The two versions of the trait are called *phenotypes*. We shall call the phenotype exhibited by individuals with genotypes AA and Aa the *dominant trait,* and the other trait will be called the *recessive trait*. In population genetics studies, it is common to have information on the phenotypes of individuals, but it is rather difficult to determine genotypes. However, some information about genotypes can be obtained by observing phenotypes of parents and children.

Assume that the allele A is dominant, that individuals mate independently of genotype, and that the genotypes AA, Aa, and aa occur in the population with probabilities 1/4, 1/2, and 1/4, respectively. We are going to observe an individual whose parents are not available, and we shall observe the phenotype of this individual. Let E be the event that the observed individual has the dominant trait. We would like to revise our opinion of the possible genotypes of the parents. There are six possible genotype combinations, $B_1, \ldots, B_6$, for the parents prior to making any observations, and these are listed in Table 2.2.

The probabilities of the B_i were computed using the assumption that the parents mated independently of genotype. For example, B_3 occurs if the father is AA and the mother is aa (probability 1/16) or if the father is aa and the mother is AA (probability 1/16). The values of $\Pr(E|B_i)$ were computed assuming that the two available alleles are passed from parents to children with probability 1/2 each and independently for the two parents. For example, given B_4, the event E occurs if and only if the child does not get two a's. The probability of getting a from both parents given B_4 is 1/4, so $\Pr(E|B_4) = 3/4$.

Now we shall compute $\Pr(B_1|E)$ and $\Pr(B_5|E)$. We leave the other calculations to the reader. The denominator of Bayes' theorem is the same for both calculations, namely,

$$\Pr(E) = \sum_{i=1}^{5}\Pr(B_i)\Pr(E|B_i)$$
$$= \frac{1}{16}\times 1 + \frac{1}{4}\times 1 + \frac{1}{8}\times 1 + \frac{1}{4}\times\frac{3}{4} + \frac{1}{4}\times\frac{1}{2} + \frac{1}{16}\times 0 = \frac{3}{4}.$$

Table 2.2 Parental genotypes for Example 2.3.5

	(AA, AA)	(AA, Aa)	(AA, aa)	(Aa, Aa)	(Aa, aa)	(aa, aa)
Name of event	B_1	B_2	B_3	B_4	B_5	B_6
Probability of B_i	1/16	1/4	1/8	1/4	1/4	1/16
$\Pr(E\|B_i)$	1	1	1	3/4	1/2	0

Applying Bayes' theorem, we get

$$\Pr(B_1|E) = \frac{\frac{1}{16} \times 1}{\frac{3}{4}} = \frac{1}{12}, \quad \Pr(B_5|E) = \frac{\frac{1}{4} \times \frac{1}{2}}{\frac{3}{4}} = \frac{1}{6}. \qquad \blacktriangleleft$$

Note: Conditional Version of Bayes' Theorem. There is also a version of Bayes' theorem conditional on an event C:

$$\Pr(B_i|A \cap C) = \frac{\Pr(B_i|C) \Pr(A|B_i \cap C)}{\sum_{j=1}^{k} \Pr(B_j|C) \Pr(A|B_j \cap C)}. \tag{2.3.2}$$

Prior and Posterior Probabilities

In Example 2.3.4, a probability like $\Pr(B_2)$ is often called the *prior probability* that the selected item will have been produced by machine M_2, because $\Pr(B_2)$ is the probability of this event before the item is selected and before it is known whether the selected item is defective or nondefective. A probability like $\Pr(B_2|A)$ is then called the *posterior probability* that the selected item was produced by machine M_2, because it is the probability of this event after it is known that the selected item is defective.

Thus, in Example 2.3.4, the prior probability that the selected item will have been produced by machine M_2 is 0.3. After an item has been selected and has been found to be defective, the posterior probability that the item was produced by machine M_2 is 0.26. Since this posterior probability is smaller than the prior probability that the item was produced by machine M_2, the posterior probability that the item was produced by one of the other machines must be larger than the prior probability that it was produced by one of those machines (see Exercises 1 and 2 at the end of this section).

Computation of Posterior Probabilities in More Than One Stage

Suppose that a box contains one fair coin and one coin with a head on each side. Suppose also that one coin is selected at random and that when it is tossed, a head is obtained. We shall determine the probability that the coin is the fair coin.

Let B_1 be the event that the coin is fair, let B_2 be the event that the coin has two heads, and let H_1 be the event that a head is obtained when the coin is tossed. Then, by Bayes' theorem,

$$\begin{aligned} \Pr(B_1|H_1) &= \frac{\Pr(B_1) \Pr(H_1|B_1)}{\Pr(B_1) \Pr(H_1|B_1) + \Pr(B_2) \Pr(H_1|B_2)} \\ &= \frac{(1/2)(1/2)}{(1/2)(1/2) + (1/2)(1)} = \frac{1}{3}. \end{aligned} \tag{2.3.3}$$

Thus, after the first toss, the posterior probability that the coin is fair is 1/3.

Now suppose that the same coin is tossed again and we assume that the two tosses are conditionally independent given both B_1 and B_2. Suppose that another head is obtained. There are two ways of determining the new value of the posterior probability that the coin is fair.

The first way is to return to the beginning of the experiment and assume again that the prior probabilities are $\Pr(B_1) = \Pr(B_2) = 1/2$. We shall let $H_1 \cap H_2$ denote the event in which heads are obtained on two tosses of the coin, and we shall calculate the posterior probability $\Pr(B_1|H_1 \cap H_2)$ that the coin is fair after we have observed the

event $H_1 \cap H_2$. The assumption that the tosses are conditionally independent given B_1 means that $\Pr(H_1 \cap H_2|B_1) = 1/2 \times 1/2 = 1/4$. By Bayes' theorem,

$$\begin{aligned}\Pr(B_1|H_1 \cap H_2) &= \frac{\Pr(B_1)\Pr(H_1 \cap H_2|B_1)}{\Pr(B_1)\Pr(H_1 \cap H_2|B_1) + \Pr(B_2)\Pr(H_1 \cap H_2|B_2)} \\ &= \frac{(1/2)(1/4)}{(1/2)(1/4) + (1/2)(1)} = \frac{1}{5}. \end{aligned} \tag{2.3.4}$$

The second way of determining this same posterior probability is to use the conditional version of Bayes' theorem (2.3.2) given the event H_1. Given H_1, the conditional probability of B_1 is 1/3, and the conditional probability of B_2 is therefore 2/3. These conditional probabilities can now serve as the prior probabilities for the next stage of the experiment, in which the coin is tossed a second time. Thus, we can apply (2.3.2) with $C = H_1$, $\Pr(B_1|H_1) = 1/3$, and $\Pr(B_2|H_1) = 2/3$. We can then compute the posterior probability $\Pr(B_1|H_1 \cap H_2)$ that the coin is fair after we have observed a head on the second toss and a head on the first toss. We shall need $\Pr(H_2|B_1 \cap H_1)$, which equals $\Pr(H_2|B_1) = 1/2$ by Theorem 2.2.4 since H_1 and H_2 are conditionally independent given B_1. Since the coin is two-headed when B_2 occurs, $\Pr(H_2|B_2 \cap H_1) = 1$. So we obtain

$$\begin{aligned}\Pr(B_1|H_1 \cap H_2) &= \frac{\Pr(B_1|H_1)\Pr(H_2|B_1 \cap H_1)}{\Pr(B_1|H_1)\Pr(H_2|B_1 \cap H_1) + \Pr(B_2|H_1)\Pr(H_2|B_2 \cap H_1)} \\ &= \frac{(1/3)(1/2)}{(1/3)(1/2) + (2/3)(1)} = \frac{1}{5}. \end{aligned} \tag{2.3.5}$$

The posterior probability of the event B_1 obtained in the second way is the same as that obtained in the first way. We can make the following general statement: If an experiment is carried out in more than one stage, then the posterior probability of every event can also be calculated in more than one stage. After each stage has been carried out, the posterior probability calculated for the event after that stage serves as the prior probability for the next stage. The reader should look back at (2.3.2) to see that this interpretation is precisely what the conditional version of Bayes' theorem says. The example we have been doing with coin tossing is typical of many applications of Bayes' theorem and its conditional version because we are assuming that the observable events are conditionally independent given each element of the partition $B_1, \ldots, B_k$ (in this case, $k = 2$). The conditional independence makes the probability of H_i (head on ith toss) given B_1 (or given B_2) the same whether or not we also condition on earlier tosses (see Theorem 2.2.4).

Conditionally Independent Events

The calculations that led to (2.3.3) and (2.3.5) together with Example 2.2.10 illustrate simple cases of a very powerful statistical model for observable events. It is very common to encounter a sequence of events that we believe are similar in that they all have the same probability of occurring. It is also common that the order in which the events are labeled does not affect the probabilities that we assign. However, we often believe that these events are not independent, because, if we were to observe some of them, we would change our minds about the probability of the ones we had not observed depending on how many of the observed events occur. For example, in the coin-tossing calculation leading up to Eq. (2.3.3), before any tosses occur, the probability of H_2 is the same as the probability of H_1, namely, the

denominator of (2.3.3), 3/4, as Theorem 2.1.4 says. However, after observing that the event H_1 occurs, the probability of H_2 is $\Pr(H_2|H_1)$, which is the denominator of (2.3.5), 5/6, as computed by the conditional version of the law of total probability (2.1.5). Even though we might treat the coin tosses as independent conditional on the coin being fair, and we might treat them as independent conditional on the coin being two-headed (in which case we know what will happen every time anyway), we cannot treat them as independent without the conditioning information. The conditioning information removes an important source of uncertainty from the problem, so we partition the sample space accordingly. Now we can use the conditional independence of the tosses to calculate joint probabilities of various combinations of events conditionally on the partition events. Finally, we can combine these probabilities using Theorem 2.1.4 and (2.1.5). Two more examples will help to illustrate these ideas.

Example 2.3.6

Learning about a Proportion. In Example 2.2.10 on page 72, a machine produced defective parts in one of two proportions, $p = 0.01$ or $p = 0.4$. Suppose that the prior probability that $p = 0.01$ is 0.9. After sampling six parts at random, suppose that we observe two defectives. What is the posterior probability that $p = 0.01$?

Let $B_1 = \{p = 0.01\}$ and $B_2 = \{p = 0.4\}$ as in Example 2.2.10. Let A be the event that two defectives occur in a random sample of size six. The prior probability of B_1 is 0.9, and the prior probability of B_2 is 0.1. We already computed $\Pr(A|B_1) = 1.44 \times 10^{-3}$ and $\Pr(A|B_2) = 0.311$ in Example 2.2.10. Bayes' theorem tells us that

$$\Pr(B_1|A) = \frac{0.9 \times 1.44 \times 10^{-3}}{0.9 \times 1.44 \times 10^{-3} + 0.1 \times 0.311} = 0.04.$$

Even though we thought originally that B_1 had probability as high as 0.9, after we learned that there were two defective items in a sample as small as six, we changed our minds dramatically and now we believe that B_1 has probability as small as 0.04. The reason for this major change is that the event A that occurred has much higher probability if B_2 is true than if B_1 is true. ◀

Example 2.3.7

A Clinical Trial. Consider the same clinical trial described in Examples 2.1.12 and 2.1.13. Let E_i be the event that the ith patient has success as her outcome. Recall that B_j is the event that $p = (j-1)/10$ for $j = 1, \ldots, 11$, where p is the proportion of successes among all possible patients. If we knew which B_j occurred, we would say that $E_1, E_2, \ldots$ were independent. That is, we are willing to model the patients as conditionally independent given each event B_j, and we set $\Pr(E_i|B_j) = (j-1)/10$ for all i, j. We shall still assume that $\Pr(B_j) = 1/11$ for all j prior to the start of the trial. We are now in position to express what we learn about p by computing posterior probabilities for the B_j events after each patient finishes the trial.

For example, consider the first patient. We calculated $\Pr(E_1) = 1/2$ in (2.1.6). If E_1 occurs, we apply Bayes' theorem to get

$$\Pr(B_j|E_1) = \frac{\Pr(E_1|B_j)\Pr(B_j)}{1/2} = \frac{2(j-1)}{10 \times 11} = \frac{j-1}{55}. \tag{2.3.6}$$

After observing one success, the posterior probabilities of large values of p are higher than their prior probabilities and the posterior probabilities of low values of p are lower than their prior probabilities as we would expect. For example, $\Pr(B_1|E_1) = 0$, because $p = 0$ is ruled out after one success. Also, $\Pr(B_2|E_1) = 0.0182$, which is much smaller than its prior value 0.0909, and $\Pr(B_{11}|E_1) = 0.1818$, which is larger than its prior value 0.0909.

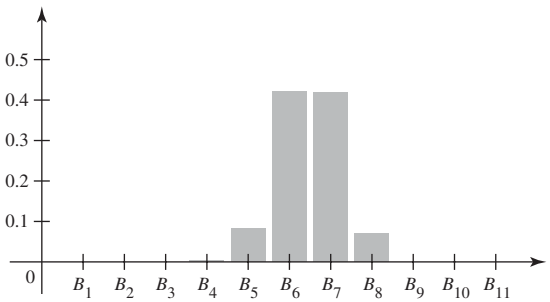

Figure 2.3 The posterior probabilities of partition elements after 40 patients in Example 2.3.7.

We could check how the posterior probabilities behave after each patient is observed. However, we shall skip ahead to the point at which all 40 patients in the imipramine column of Table 2.1 have been observed. Let A stand for the observed event that 22 of them are successes and 18 are failures. We can use the same reasoning as in Example 2.2.5 to compute $\Pr(A|B_j)$. There are $\binom{40}{22}$ possible sequences of 40 patients with 22 successes, and, conditional on B_j, the probability of each sequence is $([j-1]/10)^{22}(1-[j-1]/10)^{18}$.

So,

$$\Pr(A|B_j) = \binom{40}{22}([j-1]/10)^{22}(1-[j-1]/10)^{18}, \tag{2.3.7}$$

for each j. Then Bayes' theorem tells us that

$$\Pr(B_j|A) = \frac{\frac{1}{11}\binom{40}{22}([j-1]/10)^{22}(1-[j-1]/10)^{18}}{\sum_{i=1}^{11}\frac{1}{11}\binom{40}{22}([i-1]/10)^{22}(1-[i-1]/10)^{18}}.$$

Figure 2.3 shows the posterior probabilities of the 11 partition elements after observing A. Notice that the probabilities of B_6 and B_7 are the highest, 0.42. This corresponds to the fact that the proportion of successes in the observed sample is $22/40 = 0.55$, halfway between $(6-1)/10$ and $(7-1)/10$.

We can also compute the probability that the next patient will be a success both before the trial and after the 40 patients. Before the trial, $\Pr(E_{41}) = \Pr(E_1)$, which equals 1/2, as computed in (2.1.6). After observing the 40 patients, we can compute $\Pr(E_{41}|A)$ using the conditional version of the law of total probability, (2.1.5):

$$\Pr(E_{41}|A) = \sum_{j=1}^{11} \Pr(E_{41}|B_j \cap A)\Pr(B_j|A). \tag{2.3.8}$$

Using the values of $\Pr(B_j|A)$ in Fig. 2.3 and the fact that $\Pr(E_{41}|B_j \cap A) = \Pr(E_{41}|B_j)$ $= (j-1)/10$ (conditional independence of the E_i given the B_j), we compute (2.3.8) to be 0.5476. This is also very close to the observed frequency of success. ◀

The calculation at the end of Example 2.3.7 is typical of what happens after observing many conditionally independent events with the same conditional probability of occurrence. The conditional probability of the next event given those that were observed tends to be close to the observed frequency of occurrence among the observed events. Indeed, when there is substantial data, the choice of prior probabilities becomes far less important.

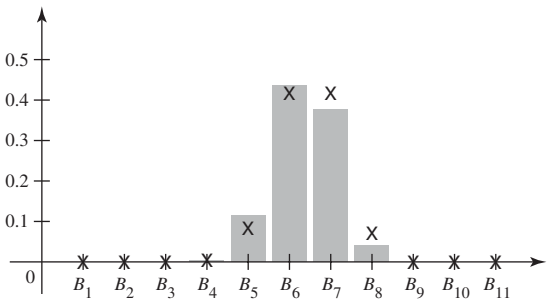

Figure 2.4 The posterior probabilities of partition elements after 40 patients in Example 2.3.8. The X characters mark the values of the posterior probabilities calculated in Example 2.3.7.

Example 2.3.8 The Effect of Prior Probabilities. Consider the same clinical trial as in Example 2.3.7. This time, suppose that a different researcher has a different prior opinion about the value of p, the probability of success. This researcher believes the following prior probabilities:

Event	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}	B_{11}
p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Prior prob.	0.00	0.19	0.19	0.17	0.14	0.11	0.09	0.06	0.04	0.01	0.00

We can recalculate the posterior probabilities using Bayes' theorem, and we get the values pictured in Fig. 2.4. To aid comparison, the posterior probabilities from Example 2.3.7 are also plotted in Fig. 2.4 using the symbol X. One can see how close the two sets of posterior probabilities are despite the large differences between the prior probabilities. If there had been fewer patients observed, there would have been larger differences between the two sets of posterior probabilites because the observed events would have provided less information. (See Exercise 12 in this section.) ◀

Summary

Bayes' theorem tells us how to compute the conditional probability of each event in a partition given an observed event A. A major use of partitions is to divide the sample space into small enough pieces so that a collection of events of interest become conditionally independent given each event in the partition.

Exercises

1. Suppose that k events $B_1, \ldots, B_k$ form a partition of the sample space S. For $i = 1, \ldots, k$, let $\Pr(B_i)$ denote the prior probability of B_i. Also, for each event A such that $\Pr(A) > 0$, let $\Pr(B_i|A)$ denote the posterior probability of B_i given that the event A has occurred. Prove that if $\Pr(B_1|A) < \Pr(B_1)$, then $\Pr(B_i|A) > \Pr(B_i)$ for at least one value of i $(i = 2, \ldots, k)$.

2. Consider again the conditions of Example 2.3.4 in this section, in which an item was selected at random from a batch of manufactured items and was found to be defective. For which values of i ($i = 1, 2, 3$) is the posterior probability that the item was produced by machine M_i larger than the prior probability that the item was produced by machine M_i?

3. Suppose that in Example 2.3.4 in this section, the item selected at random from the entire lot is found to be nondefective. Determine the posterior probability that it was produced by machine M_2.

4. A new test has been devised for detecting a particular type of cancer. If the test is applied to a person who has this type of cancer, the probability that the person will have a positive reaction is 0.95 and the probability that the person will have a negative reaction is 0.05. If the test is applied to a person who does not have this type of cancer, the probability that the person will have a positive reaction is 0.05 and the probability that the person will have a negative reaction is 0.95. Suppose that in the general population, one person out of every 100,000 people has this type of cancer. If a person selected at random has a positive reaction to the test, what is the probability that he has this type of cancer?

5. In a certain city, 30 percent of the people are Conservatives, 50 percent are Liberals, and 20 percent are Independents. Records show that in a particular election, 65 percent of the Conservatives voted, 82 percent of the Liberals voted, and 50 percent of the Independents voted. If a person in the city is selected at random and it is learned that she did not vote in the last election, what is the probability that she is a Liberal?

6. Suppose that when a machine is adjusted properly, 50 percent of the items produced by it are of high quality and the other 50 percent are of medium quality. Suppose, however, that the machine is improperly adjusted during 10 percent of the time and that, under these conditions, 25 percent of the items produced by it are of high quality and 75 percent are of medium quality.

- **a.** Suppose that five items produced by the machine at a certain time are selected at random and inspected. If four of these items are of high quality and one item is of medium quality, what is the probability that the machine was adjusted properly at that time?
- **b.** Suppose that one additional item, which was produced by the machine at the same time as the other five items, is selected and found to be of medium quality. What is the new posterior probability that the machine was adjusted properly?

7. Suppose that a box contains five coins and that for each coin there is a different probability that a head will be obtained when the coin is tossed. Let p_i denote the probability of a head when the ith coin is tossed ($i = 1, \ldots, 5$), and suppose that $p_1 = 0$, $p_2 = 1/4$, $p_3 = 1/2$, $p_4 = 3/4$, and $p_5 = 1$.

- **a.** Suppose that one coin is selected at random from the box and when it is tossed once, a head is obtained. What is the posterior probability that the ith coin was selected ($i = 1, \ldots, 5$)?
- **b.** If the same coin were tossed again, what would be the probability of obtaining another head?
- **c.** If a tail had been obtained on the first toss of the selected coin and the same coin were tossed again, what would be the probability of obtaining a head on the second toss?

8. Consider again the box containing the five different coins described in Exercise 7. Suppose that one coin is selected at random from the box and is tossed repeatedly until a head is obtained.

- **a.** If the first head is obtained on the fourth toss, what is the posterior probability that the ith coin was selected ($i = 1, \ldots, 5$)?
- **b.** If we continue to toss the same coin until another head is obtained, what is the probability that exactly three additional tosses will be required?

9. Consider again the conditions of Exercise 14 in Sec. 2.1. Suppose that several parts will be observed and that the different parts are conditionally independent given each of the three states of repair of the machine. If seven parts are observed and exactly one is defective, compute the posterior probabilities of the three states of repair.

10. Consider again the conditions of Example 2.3.5, in which the phenotype of an individual was observed and found to be the dominant trait. For which values of i ($i = 1, \ldots, 6$) is the posterior probability that the parents have the genotypes of event B_i smaller than the prior probability that the parents have the genotyes of event B_i?

11. Suppose that in Example 2.3.5 the observed individual has the recessive trait. Determine the posterior probability that the parents have the genotypes of event B_4.

12. In the clinical trial in Examples 2.3.7 and 2.3.8, suppose that we have only observed the first five patients and three of the five had been successes. Use the two different sets of prior probabilities from Examples 2.3.7 and 2.3.8 to calculate two sets of posterior probabilities. Are these two sets of posterior probabilities as close to each other as were the two in Examples 2.3.7 and 2.3.8? Why or why not?

13. Suppose that a box contains one fair coin and one coin with a head on each side. Suppose that a coin is drawn at random from this box and that we begin to flip the coin. In Eqs. (2.3.4) and (2.3.5), we computed the conditional

probability that the coin was fair given that the first two flips both produce heads.

a. Suppose that the coin is flipped a third time and another head is obtained. Compute the probability that the coin is fair given that all three flips produced heads.

b. Suppose that the coin is flipped a fourth time and the result is tails. Compute the posterior probability that the coin is fair.

14. Consider again the conditions of Exercise 23 in Sec. 2.2. Assume that $\Pr(B) = 0.4$. Let A be the event that exactly 8 out of 11 programs compiled. Compute the conditional probability of B given A.

15. Use the prior probabilities in Example 2.3.8 for the events $B_1, \ldots, B_{11}$. Let E_1 be the event that the first patient is a success. Compute the probability of E_1 and explain why it is so much less than the value computed in Example 2.3.7.

16. Consider a machine that produces items in sequence. Under normal operating conditions, the items are independent with probability 0.01 of being defective. However, it is possible for the machine to develop a "memory" in the following sense: After each defective item, and independent of anything that happened earlier, the probability that the next item is defective is 2/5. After each nondefective item, and independent of anything that happened earlier, the probability that the next item is defective is 1/165.

Assume that the machine is either operating normally for the whole time we observe or has a memory for the whole time that we observe. Let B be the event that the machine is operating normally, and assume that $\Pr(B) = 2/3$. Let D_i be the event that the ith item inspected is defective. Assume that D_1 is independent of B.

a. Prove that $\Pr(D_i) = 0.01$ for all i. *Hint:* Use induction.

b. Assume that we observe the first six items and the event that occurs is $E = D_1^c \cap D_2^c \cap D_3 \cap D_4 \cap D_5^c \cap D_6^c$. That is, the third and fourth items are defective, but the other four are not. Compute $\Pr(B|D)$.

⋆ 2.4 The Gambler's Ruin Problem

Consider two gamblers with finite resources who repeatedly play the same game against each other. Using the tools of conditional probability, we can calculate the probability that each of the gamblers will eventually lose all of his money to the opponent.

Statement of the Problem

Suppose that two gamblers A and B are playing a game against each other. Let p be a given number ($0 < p < 1$), and suppose that on each play of the game, the probability that gambler A will win one dollar from gambler B is p and the probability that gambler B will win one dollar from gambler A is $1 - p$. Suppose also that the initial fortune of gambler A is i dollars and the initial fortune of gambler B is $k - i$ dollars, where i and $k - i$ are given positive integers. Thus, the total fortune of the two gamblers is k dollars. Finally, suppose that the gamblers play the game repeatedly and independently until the fortune of one of them has been reduced to 0 dollars. Another way to think about this problem is that B is a casino and A is a gambler who is determined to quit as soon he wins $k - i$ dollars from the casino or when he goes broke, whichever comes first.

We shall now consider this game from the point of view of gambler A. His initial fortune is i dollars and on each play of the game his fortune will either increase by one dollar with a probability of p or decrease by one dollar with a probability of $1 - p$. If $p > 1/2$, the game is favorable to him; if $p < 1/2$, the game is unfavorable to him; and if $p = 1/2$, the game is equally favorable to both gamblers. The game ends either when the fortune of gambler A reaches k dollars, in which case gambler B will have no money left, or when the fortune of gambler A reaches 0 dollars. The problem is to

determine the probability that the fortune of gambler A will reach k dollars before it reaches 0 dollars. Because one of the gamblers will have no money left at the end of the game, this problem is called the *Gambler's Ruin* problem.

Solution of the Problem

We shall continue to assume that the total fortune of the gamblers A and B is k dollars, and we shall let a_i denote the probability that the fortune of gambler A will reach k dollars before it reaches 0 dollars, given that his initial fortune is i dollars. We assume that the game is the same each time it is played and the plays are independent of each other. It follows that, after each play, the Gambler's Ruin problem essentially starts over with the only change being that the initial fortunes of the two gamblers have changed. In particular, for each $j = 0, \ldots, k$, each time that we observe a sequence of plays that lead to gambler A's fortune being j dollars, the conditional probability, given such a sequence, that gambler A wins is a_j. If gambler A's fortune ever reaches 0, then gambler A is ruined, hence $a_0 = 0$. Similarly, if his fortune ever reaches k, then gambler A has won, hence $a_k = 1$. We shall now determine the value of a_i for $i = 1, \ldots, k-1$.

Let A_1 denote the event that gambler A wins one dollar on the first play of the game, let B_1 denote the event that gambler A loses one dollar on the first play of the game, and let W denote the event that the fortune of gambler A ultimately reaches k dollars before it reaches 0 dollars. Then

$$\begin{aligned} \Pr(W) &= \Pr(A_1)\Pr(W|A_1) + \Pr(B_1)\Pr(W|B_1) \\ &= p\Pr(W|A_1) + (1-p)\Pr(W|B_1). \end{aligned} \tag{2.4.1}$$

Since the initial fortune of gambler A is i dollars ($i = 1, \ldots, k-1$), then $\Pr(W) = a_i$. Furthermore, if gambler A wins one dollar on the first play of the game, then his fortune becomes $i+1$ dollars and the conditional probability $\Pr(W|A_1)$ that his fortune will ultimately reach k dollars is therefore a_{i+1}. If A loses one dollar on the first play of the game, then his fortune becomes $i-1$ dollars and the conditional probability $\Pr(W|B_1)$ that his fortune will ultimately reach k dollars is therefore a_{i-1}. Hence, by Eq. (2.4.1),

$$a_i = pa_{i+1} + (1-p)a_{i-1}. \tag{2.4.2}$$

We shall let $i = 1, \ldots, k-1$ in Eq. (2.4.2). Then, since $a_0 = 0$ and $a_k = 1$, we obtain the following $k-1$ equations:

$$\begin{aligned} a_1 &= pa_2, \\ a_2 &= pa_3 + (1-p)a_1, \\ a_3 &= pa_4 + (1-p)a_2, \\ &\vdots \\ a_{k-2} &= pa_{k-1} + (1-p)a_{k-3}, \\ a_{k-1} &= p + (1-p)a_{k-2}. \end{aligned} \tag{2.4.3}$$

If the value of a_i on the left side of the ith equation is rewritten in the form $pa_i + (1-p)a_i$ and some elementary algebra is performed, then these $k-1$ equations can

be rewritten as follows:

$$\begin{aligned}
a_2 - a_1 &= \frac{1-p}{p} a_1, \\
a_3 - a_2 &= \frac{1-p}{p}(a_2 - a_1) &&= \left(\frac{1-p}{p}\right)^2 a_1, \\
a_4 - a_3 &= \frac{1-p}{p}(a_3 - a_2) &&= \left(\frac{1-p}{p}\right)^3 a_1, \\
&\vdots \\
a_{k-1} - a_{k-2} &= \frac{1-p}{p}(a_{k-2} - a_{k-3}) = \left(\frac{1-p}{p}\right)^{k-2} a_1, \\
1 - a_{k-1} &= \frac{1-p}{p}(a_{k-1} - a_{k-2}) = \left(\frac{1-p}{p}\right)^{k-1} a_1.
\end{aligned} \tag{2.4.4}$$

By equating the sum of the left sides of these $k-1$ equations with the sum of the right sides, we obtain the relation

$$1 - a_1 = a_1 \sum_{i=1}^{k-1} \left(\frac{1-p}{p}\right)^i. \tag{2.4.5}$$

Solution for a Fair Game Suppose first that $p = 1/2$. Then $(1-p)/p = 1$, and it follows from Eq. (2.4.5) that $1 - a_1 = (k-1)a_1$, from which $a_1 = 1/k$. In turn, it follows from the first equation in (2.4.4) that $a_2 = 2/k$, it follows from the second equation in (2.4.4) that $a_3 = 3/k$, and so on. In this way, we obtain the following complete solution when $p = 1/2$:

$$a_i = \frac{i}{k} \quad \text{for } i = 1, \ldots, k-1. \tag{2.4.6}$$

Example 2.4.1 The Probability of Winning in a Fair Game. Suppose that $p = 1/2$, in which case the game is equally favorable to both gamblers; and suppose that the initial fortune of gambler A is 98 dollars and the initial fortune of gambler B is just two dollars. In this example, $i = 98$ and $k = 100$. Therefore, it follows from Eq. (2.4.6) that there is a probability of 0.98 that gambler A will win two dollars from gambler B before gambler B wins 98 dollars from gambler A. ◀

Solution for an Unfair Game Suppose now that $p \neq 1/2$. Then Eq. (2.4.5) can be rewritten in the form

$$1 - a_1 = a_1 \frac{\left(\frac{1-p}{p}\right)^k - \left(\frac{1-p}{p}\right)}{\left(\frac{1-p}{p}\right) - 1}. \tag{2.4.7}$$

Hence,

$$a_1 = \frac{\left(\frac{1-p}{p}\right) - 1}{\left(\frac{1-p}{p}\right)^k - 1}. \tag{2.4.8}$$

Each of the other values of a_i for $i = 2, \ldots, k-1$ can now be determined in turn from the equations in (2.4.4). In this way, we obtain the following complete solution:

$$a_i = \frac{\left(\frac{1-p}{p}\right)^i - 1}{\left(\frac{1-p}{p}\right)^k - 1} \quad \text{for } i = 1, \ldots, k-1. \tag{2.4.9}$$

Example 2.4.2

The Probability of Winning in an Unfavorable Game. Suppose that $p = 0.4$, in which case the probability that gambler A will win one dollar on any given play is smaller than the probability that he will lose one dollar. Suppose also that the initial fortune of gambler A is 99 dollars and the initial fortune of gambler B is just one dollar. We shall determine the probability that gambler A will win one dollar from gambler B before gambler B wins 99 dollars from gambler A.

In this example, the required probability a_i is given by Eq. (2.4.9), in which $(1-p)/p = 3/2$, $i = 99$, and $k = 100$. Therefore,

$$a_i = \frac{\left(\frac{3}{2}\right)^{99} - 1}{\left(\frac{3}{2}\right)^{100} - 1} \approx \frac{1}{3/2} = \frac{2}{3}.$$

Hence, although the probability that gambler A will win one dollar on any given play is only 0.4, the probability that he will win one dollar before he loses 99 dollars is approximately 2/3. ◀

Summary

We considered a gambler and an opponent who each start with finite amounts of money. The two then play a sequence of games against each other until one of them runs out of money. We were able to calculate the probability that each of them would be the first to run out as a function of the probability of winning the game and of how much money each has at the start.

Exercises

1. Consider the unfavorable game in Example 2.4.2. This time, suppose that the initial fortune of gambler A is i dollars with $i \leq 98$. Suppose that the initial fortune of gambler B is $100 - i$ dollars. Show that the probability is greater than 1/2 that gambler A losses i dollars before winning $100 - i$ dollars.

2. Consider the following three different possible conditions in the gambler's ruin problem:

a. The initial fortune of gambler A is two dollars, and the initial fortune of gambler B is one dollar.

b. The initial fortune of gambler A is 20 dollars, and the initial fortune of gambler B is 10 dollars.

c. The initial fortune of gambler A is 200 dollars, and the initial fortune of gambler B is 100 dollars.

Suppose that $p = 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

3. Consider again the three different conditions (a), (b), and (c) given in Exercise 2, but suppose now that $p < 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

4. Consider again the three different conditions (a), (b), and (c) given in Exercise 2, but suppose now that $p > 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

5. Suppose that on each play of a certain game, a person is equally likely to win one dollar or lose one dollar. Suppose also that the person's goal is to win two dollars by playing this game. How large an initial fortune must the person have in order for the probability to be at least 0.99 that she will achieve her goal before she loses her initial fortune?

6. Suppose that on each play of a certain game, a person will either win one dollar with probability 2/3 or lose one dollar with probability 1/3. Suppose also that the person's goal is to win two dollars by playing this game. How large an initial fortune must the person have in order for the probability to be at least 0.99 that he will achieve his goal before he loses his initial fortune?

7. Suppose that on each play of a certain game, a person will either win one dollar with probability 1/3 or lose one dollar with probability 2/3. Suppose also that the person's goal is to win two dollars by playing this game. Show that no matter how large the person's initial fortune might be, the probability that she will achieve her goal before she loses her initial fortune is less than 1/4.

8. Suppose that the probability of a head on any toss of a certain coin is p $(0 < p < 1)$, and suppose that the coin is tossed repeatedly. Let X_n denote the total number of heads that have been obtained on the first n tosses, and let $Y_n = n - X_n$ denote the total number of tails on the first n tosses. Suppose that the tosses are stopped as soon as a number n is reached such that either $X_n = Y_n + 3$ or $Y_n = X_n + 3$. Determine the probability that $X_n = Y_n + 3$ when the tosses are stopped.

9. Suppose that a certain box A contains five balls and another box B contains 10 balls. One of these two boxes is selected at random, and one ball from the selected box is transferred to the other box. If this process of selecting a box at random and transferring one ball from that box to the other box is repeated indefinitely, what is the probability that box A will become empty before box B becomes empty?

2.5 Supplementary Exercises

1. Suppose that A, B, and D are any three events such that $\Pr(A|D) \geq \Pr(B|D)$ and $\Pr(A|D^c) \geq \Pr(B|D^c)$. Prove that $\Pr(A) \geq \Pr(B)$.

2. Suppose that a fair coin is tossed repeatedly and independently until both a head and a tail have appeared at least once. **(a)** Describe the sample space of this experiment. **(b)** What is the probability that exactly three tosses will be required?

3. Suppose that A and B are events such that $\Pr(A) = 1/3$, $\Pr(B) = 1/5$, and $\Pr(A|B) + \Pr(B|A) = 2/3$. Evaluate $\Pr(A^c \cup B^c)$.

4. Suppose that A and B are independent events such that $\Pr(A) = 1/3$ and $\Pr(B) > 0$. What is the value of $\Pr(A \cup B^c|B)$?

5. Suppose that in 10 rolls of a balanced die, the number 6 appeared exactly three times. What is the probability that the first three rolls each yielded the number 6?

6. Suppose that A, B, and D are events such that A and B are independent, $\Pr(A \cap B \cap D) = 0.04$, $\Pr(D|A \cap B) = 0.25$, and $\Pr(B) = 4\Pr(A)$. Evaluate $\Pr(A \cup B)$.

7. Suppose that the events A, B, and C are mutually independent. Under what conditions are A^c, B^c, and C^c mutually independent?

8. Suppose that the events A and B are disjoint and that each has positive probability. Are A and B independent?

9. Suppose that A, B, and C are three events such that A and B are disjoint, A and C are independent, and B and C are independent. Suppose also that $4\Pr(A) = 2\Pr(B) = \Pr(C) > 0$ and $\Pr(A \cup B \cup C) = 5\Pr(A)$. Determine the value of $\Pr(A)$.

10. Suppose that each of two dice is loaded so that when either die is rolled, the probability that the number k will appear is 0.1 for $k = 1, 2, 5$, or 6 and is 0.3 for $k = 3$ or 4. If the two loaded dice are rolled independently, what is the probability that the sum of the two numbers that appear will be 7?

11. Suppose that there is a probability of 1/50 that you will win a certain game. If you play the game 50 times, independently, what is the probability that you will win at least once?

12. Suppose that a balanced die is rolled three times, and let X_i denote the number that appears on the ith roll $(i = 1, 2, 3)$. Evaluate $\Pr(X_1 > X_2 > X_3)$.

13. Three students A, B, and C are enrolled in the same class. Suppose that A attends class 30 percent of the time, B attends class 50 percent of the time, and C attends class 80 percent of the time. If these students attend class independently of each other, what is (a) the probability that at least one of them will be in class on a particular day and (b) the probability that exactly one of them will be in class on a particular day?

14. Consider the World Series of baseball, as described in Exercise 16 of Sec. 2.2. If there is probability p that team A will win any particular game, what is the probability

that it will be necessary to play seven games in order to determine the winner of the Series?

15. Suppose that three red balls and three white balls are thrown at random into three boxes and and that all throws are independent. What is the probability that each box contains one red ball and one white ball?

16. If five balls are thrown at random into n boxes, and all throws are independent, what is the probability that no box contains more than two balls?

17. Bus tickets in a certain city contain four numbers, U, V, W, and X. Each of these numbers is equally likely to be any of the 10 digits 0, 1, . . . , 9, and the four numbers are chosen independently. A bus rider is said to be lucky if $U + V = W + X$. What proportion of the riders are lucky?

18. A certain group has eight members. In January, three members are selected at random to serve on a committee. In February, four members are selected at random and independently of the first selection to serve on another committee. In March, five members are selected at random and independently of the previous two selections to serve on a third committee. Determine the probability that each of the eight members serves on at least one of the three committees.

19. For the conditions of Exercise 18, determine the probability that two particular members A and B will serve together on at least one of the three committees.

20. Suppose that two players A and B take turns rolling a pair of balanced dice and that the winner is the first player who obtains the sum of 7 on a given roll of the two dice. If A rolls first, what is the probability that B will win?

21. Three players A, B, and C take turns tossing a fair coin. Suppose that A tosses the coin first, B tosses second, and C tosses third; and suppose that this cycle is repeated indefinitely until someone wins by being the first player to obtain a head. Determine the probability that each of three players will win.

22. Suppose that a balanced die is rolled repeatedly until the same number appears on two successive rolls, and let X denote the number of rolls that are required. Determine the value of $\Pr(X = x)$, for $x = 2, 3, \ldots$.

23. Suppose that 80 percent of all statisticians are shy, whereas only 15 percent of all economists are shy. Suppose also that 90 percent of the people at a large gathering are economists and the other 10 percent are statisticians. If you meet a shy person at random at the gathering, what is the probability that the person is a statistician?

24. Dreamboat cars are produced at three different factories A, B, and C. Factory A produces 20 percent of the total output of Dreamboats, B produces 50 percent, and C produces 30 percent. However, 5 percent of the cars produced at A are lemons, 2 percent of those produced at B are lemons, and 10 percent of those produced at C are lemons. If you buy a Dreamboat and it turns out to be a lemon, what is the probability that it was produced at factory A?

25. Suppose that 30 percent of the bottles produced in a certain plant are defective. If a bottle is defective, the probability is 0.9 that an inspector will notice it and remove it from the filling line. If a bottle is not defective, the probability is 0.2 that the inspector will think that it is defective and remove it from the filling line.

a. If a bottle is removed from the filling line, what is the probability that it is defective?

b. If a customer buys a bottle that has not been removed from the filling line, what is the probability that it is defective?

26. Suppose that a fair coin is tossed until a head is obtained and that this entire experiment is then performed independently a second time. What is the probability that the second experiment requires more tosses than the first experiment?

27. Suppose that a family has exactly n children ($n \geq 2$). Assume that the probability that any child will be a girl is 1/2 and that all births are independent. Given that the family has at least one girl, determine the probability that the family has at least one boy.

28. Suppose that a fair coin is tossed independently n times. Determine the probability of obtaining exactly $n - 1$ heads, given **(a)** that at least $n - 2$ heads are obtained and **(b)** that heads are obtained on the first $n - 2$ tosses.

29. Suppose that 13 cards are selected at random from a regular deck of 52 playing cards.

a. If it is known that at least one ace has been selected, what is the probability that at least two aces have been selected?

b. If it is known that the ace of hearts has been selected, what is the probability that at least two aces have been selected?

30. Suppose that n letters are placed at random in n envelopes, as in the matching problem of Sec. 1.10, and let q_n denote the probability that no letter is placed in the correct envelope. Show that the probability that exactly one letter is placed in the correct envelope is q_{n-1}.

31. Consider again the conditions of Exercise 30. Show that the probability that exactly two letters are placed in the correct envelopes is $(1/2)q_{n-2}$.

32. Consider again the conditions of Exercise 7 of Sec. 2.2. If exactly one of the two students A and B is in class on a given day, what is the probability that it is A?

33. Consider again the conditions of Exercise 2 of Sec. 1.10. If a family selected at random from the city

subscribes to exactly one of the three newspapers A, B, and C, what is the probability that it is A?

34. Three prisoners A, B, and C on death row know that exactly two of them are going to be executed, but they do not know which two. Prisoner A knows that the jailer will not tell him whether or not he is going to be executed. He therefore asks the jailer to tell him the name of one prisoner other than A himself who will be executed. The jailer responds that B will be executed. Upon receiving this response, Prisoner A reasons as follows: Before he spoke to the jailer, the probability was 2/3 that he would be one of the two prisoners executed. After speaking to the jailer, he knows that either he or prisoner C will be the other one to be executed. Hence, the probability that he will be executed is now only 1/2. Thus, merely by asking the jailer his question, the prisoner reduced the probability that he would be executed from 2/3 to 1/2, because he could go through exactly this same reasoning regardless of which answer the jailer gave. Discuss what is wrong with prisoner A's reasoning.

35. Suppose that each of two gamblers A and B has an initial fortune of 50 dollars, and that there is probability p that gambler A will win on any single play of a game against gambler B. Also, suppose either that one gambler can win one dollar from the other on each play of the game or that they can double the stakes and one can win two dollars from the other on each play of the game. Under which of these two conditions does A have the greater probability of winning the initial fortune of B before losing her own for each of the following conditions: **(a)** $p < 1/2$; **(b)** $p > 1/2$; **(c)** $p = 1/2$?

36. A sequence of n job candidates is prepared to interview for a job. We would like to hire the best candidate, but we have no information to distinguish the candidates before we interview them. We assume that the best candidate is equally likely to be each of the n candidates in the sequence before the interviews start. After the interviews start, we are able to rank those candidates we have seen, but we have no information about where the remaining candidates rank relative to those we have seen. After each interview, it is required that either we hire the current candidate immediately and stop the interviews, or we must let the current candidate go and we never can call them back. We choose to interview as follows: We select a number $0 \le r < n$ and we interview the first r candidates without any intention of hiring them. Starting with the next candidate $r + 1$, we continue interviewing until the current candidate is the best we have seen so far. We then stop and hire the current candidate. If none of the candidates from $r + 1$ to n is the best, we just hire candidate n. We would like to compute the probability that we hire the best candidate and we would like to choose r to make this probability as large as possible. Let A be the event that we hire the best candidate, and let B_i be the event that the best candidate is in position i in the sequence of interviews.

- **a.** Let $i > r$. Find the probability that the candidate who is relatively the best among the first i interviewed appears in the first r interviews.
- **b.** Prove that $\Pr(A|B_i) = 0$ for $i \le r$ and $\Pr(A|B_i) = r/(i-1)$ for $i > r$.
- **c.** For fixed r, let p_r be the probability of A using that value of r. Prove that $p_r = (r/n) \sum_{i=r+1}^{n} (i-1)^{-1}$.
- **d.** Let $q_r = p_r - p_{r-1}$ for $r = 1, \ldots, n-1$, and prove that q_r is a strictly decreasing function of r.
- **e.** Show that a value of r that maximizes p_r is the last r such that $q_r > 0$. (*Hint:* Write $p_r = p_0 + q_1 + \cdots + q_r$ for $r > 0$.)
- **f.** For $n = 10$, find the value of r that maximizes p_r, and find the corresponding p_r value.

Chapter

3 Random Variables and Distributions

3.1 Random Variables and Discrete Distributions

A random variable is a real-valued function defined on a sample space. Random variables are the main tools used for modeling unknown quantities in statistical analyses. For each random variable X and each set C of real numbers, we could calculate the probability that X takes its value in C. The collection of all of these probabilities is the distribution of X. There are two major classes of distributions and random variables: discrete (this section) and continuous (Sec. 3.2). Discrete distributions are those that assign positive probability to at most countably many different values. A discrete distribution can be characterized by its probability function (p.f.), which specifies the probability that the random variable takes each of the different possible values. A random variable with a discrete distribution will be called a discrete random variable.

Definition of a Random Variable

Example 3.1.1 Tossing a Coin. Consider an experiment in which a fair coin is tossed 10 times. In this experiment, the sample space S can be regarded as the set of outcomes consisting of the 2^{10} different sequences of 10 heads and/or tails that are possible. We might be interested in the number of heads in the observed outcome. We can let X stand for the real-valued function defined on S that counts the number of heads in each outcome. For example, if s is the sequence HHTTTHTTTH, then $X(s) = 4$. For each possible sequence s consisting of 10 heads and/or tails, the value $X(s)$ equals the number of heads in the sequence. The possible values for the function X are $0, 1, \ldots, 10$. ◀

Definition 3.1.1 Random Variable. Let S be the sample space for an experiment. A real-valued function that is defined on S is called a *random variable*.

For example, in Example 3.1.1, the number X of heads in the 10 tosses is a random variable. Another random variable in that example is $Y = 10 - X$, the number of tails.

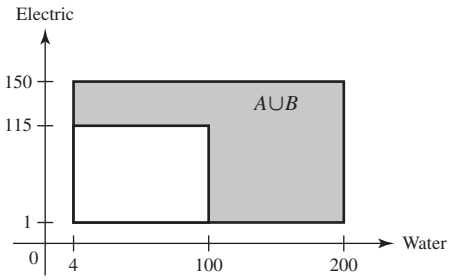

Figure 3.1 The event that at least one utility demand is high in Example 3.1.3.

Example 3.1.2 Measuring a Person's Height. Consider an experiment in which a person is selected at random from some population and her height in inches is measured. This height is a random variable. ◀

Example 3.1.3 Demands for Utilities. Consider the contractor in Example 1.5.4 on page 19 who is concerned about the demands for water and electricity in a new office complex. The sample space was pictured in Fig. 1.5 on page 12, and it consists of a collection of points of the form (x, y), where x is the demand for water and y is the demand for electricity. That is, each point $s \in S$ is a pair $s = (x, y)$. One random variable that is of interest in this problem is the demand for water. This can be expressed as $X(s) = x$ when $s = (x, y)$. The possible values of X are the numbers in the interval $[4, 200]$. Another interesting random variable is Y, equal to the electricity demand, which can be expressed as $Y(s) = y$ when $s = (x, y)$. The possible values of Y are the numbers in the interval $[1, 150]$. A third possible random variable Z is an indicator of whether or not at least one demand is high. Let A and B be the two events described in Example 1.5.4. That is, A is the event that water demand is at least 100, and B is the event that electric demand is at least 115. Define

$$Z(s) = \begin{cases} 1 & \text{if } s \in A \cup B, \\ 0 & \text{if } s \notin A \cup B. \end{cases}$$

The possible values of Z are the numbers 0 and 1. The event $A \cup B$ is indicated in Fig. 3.1. ◀

The Distribution of a Random Variable

When a probability measure has been specified on the sample space of an experiment, we can determine probabilities associated with the possible values of each random variable X. Let C be a subset of the real line such that $\{X \in C\}$ is an event, and let $\Pr(X \in C)$ denote the probability that the value of X will belong to the subset C. Then $\Pr(X \in C)$ is equal to the probability that the outcome s of the experiment will be such that $X(s) \in C$. In symbols,

$$\Pr(X \in C) = \Pr(\{s: X(s) \in C\}). \tag{3.1.1}$$

Definition 3.1.2 Distribution. Let X be a random variable. The *distribution* of X is the collection of all probabilities of the form $\Pr(X \in C)$ for all sets C of real numbers such that $\{X \in C\}$ is an event.

It is a straightforward consequence of the definition of the distribution of X that this distribution is itself a probability measure on the set of real numbers. The set

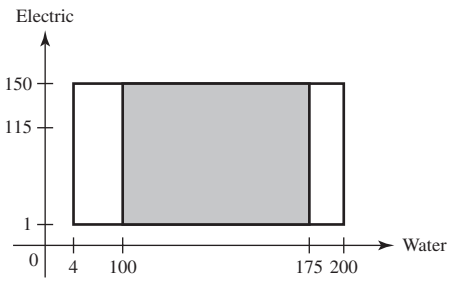

Figure 3.2 The event that water demand is between 50 and 175 in Example 3.1.5.

$\{X \in C\}$ will be an event for every set C of real numbers that most readers will be able to imagine.

Example 3.1.4 Tossing a Coin. Consider again an experiment in which a fair coin is tossed 10 times, and let X be the number of heads that are obtained. In this experiment, the possible values of X are 0, 1, 2, . . . , 10. For each x, $\Pr(X = x)$ is the sum of the probabilities of all of the outcomes in the event $\{X = x\}$. Because the coin is fair, each outcome has the same probability $1/2^{10}$, and we need only count how many outcomes s have $X(s) = x$. We know that $X(s) = x$ if and only if exactly x of the 10 tosses are H. Hence, the number of outcomes s with $X(s) = x$ is the same as the number of subsets of size x (to be the heads) that can be chosen from the 10 tosses, namely, $\binom{10}{x}$, according to Definitions 1.8.1 and 1.8.2. Hence,

$$\Pr(X = x) = \binom{10}{x}\frac{1}{2^{10}} \quad \text{for } x = 0, 1, 2, \ldots, 10. \qquad \blacktriangleleft$$

Example 3.1.5 Demands for Utilities. In Example 1.5.4, we actually calculated some features of the distributions of the three random variables X, Y, and Z defined in Example 3.1.3. For example, the event A, defined as the event that water demand is at least 100, can be expressed as $A = \{X \geq 100\}$, and $\Pr(A) = 0.5102$. This means that $\Pr(X \geq 100) = 0.5102$. The distribution of X consists of all probabilities of the form $\Pr(X \in C)$ for all sets C such that $\{X \in C\}$ is an event. These can all be calculated in a manner similar to the calculation of $\Pr(A)$ in Example 1.5.4. In particular, if C is a subinterval of the interval [4, 200], then

$$\Pr(X \in C) = \frac{(150 - 1) \times (\text{length of interval } C)}{29{,}204}. \tag{3.1.2}$$

For example, if C is the interval [50,175], then its length is 125, and $\Pr(X \in C) = 149 \times 125/29{,}204 = 0.6378$. The subset of the sample space whose probability was just calculated is drawn in Fig. 3.2. $\blacktriangleleft$

The general definition of distribution in Definition 3.1.2 is awkward, and it will be useful to find alternative ways to specify the distributions of random variables. In the remainder of this section, we shall introduce a few such alternatives.

Discrete Distributions

Definition 3.1.3 Discrete Distribution/Random Variable. We say that a random variable X has a *discrete distribution* or that X is a *discrete random variable* if X can take only a finite number k of different values $x_1, \ldots, x_k$ or, at most, an infinite sequence of different values $x_1, x_2, \ldots$.

Random variables that can take every value in an interval are said to have *continuous distributions* and are discussed in Sec. 3.2.

Definition 3.1.4 Probability Function/p.f./Support. If a random variable X has a discrete distribution, the *probability function* (abbreviated *p.f.*) of X is defined as the function f such that for every real number x,

$$f(x) = \Pr(X = x).$$

The closure of the set $\{x : f(x) > 0\}$ is called the *support of (the distribution of) X*.

Some authors refer to the probability function as the *probability mass function,* or p.m.f. We will not use that term again in this text.

Example 3.1.6 Demands for Utilities. The random variable Z in Example 3.1.3 equals 1 if at least one of the utility demands is high, and $Z = 0$ if neither demand is high. Since Z takes only two different values, it has a discrete distribution. Note that $\{s : Z(s) = 1\} = A \cup B$, where A and B are defined in Example 1.5.4. We calculated $\Pr(A \cup B) = 0.65253$ in Example 1.5.4. If Z has p.f. f, then

$$f(z) = \begin{cases} 0.65253 & \text{if } z = 1, \\ 0.34747 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The support of Z is the set $\{0, 1\}$, which has only two elements. ◀

Example 3.1.7 Tossing a Coin. The random variable X in Example 3.1.4 has only 11 different possible values. Its p.f. f is given at the end of that example for the values $x = 0, \ldots, 10$ that constitute the support of X; $f(x) = 0$ for all other values of x. ◀

Here are some simple facts about probability functions

Theorem 3.1.1 Let X be a discrete random variable with p.f. f. If x is not one of the possible values of X, then $f(x) = 0$. Also, if the sequence $x_1, x_2, \ldots$ includes all the possible values of X, then $\sum_{i=1}^{\infty} f(x_i) = 1$. ■

A typical p.f. is sketched in Fig. 3.3, in which each vertical segment represents the value of $f(x)$ corresponding to a possible value x. The sum of the heights of the vertical segments in Fig. 3.3 must be 1.

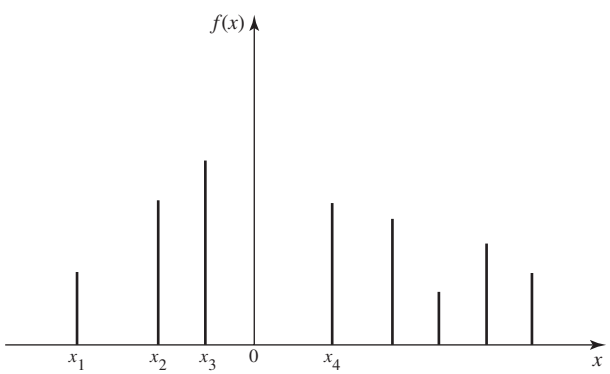

Figure 3.3 An example of a p.f.

Theorem 3.1.2 shows that the p.f. of a discrete random variable characterizes its distribution, and it allows us to dispense with the general definition of distribution when we are discussing discrete random variables.

Theorem 3.1.2

If X has a discrete distribution, the probability of each subset C of the real line can be determined from the relation

$$\Pr(X \in C) = \sum_{x_i \in C} f(x_i). \qquad \blacksquare$$

Some random variables have distributions that appear so frequently that the distributions are given names. The random variable Z in Example 3.1.6 is one such.

Definition 3.1.5

Bernoulli Distribution/Random Variable. A random variable Z that takes only two values 0 and 1 with $\Pr(Z = 1) = p$ has the *Bernoulli distribution with parameter* p. We also say that Z is a *Bernoulli random variable with parameter* p.

The Z in Example 3.1.6 has the Bernoulli distribution with parameter 0.65252. It is easy to see that the name of each Bernoulli distribution is enough to allow us to compute the p.f., which, in turn, allows us to characterize its distribution.

We conclude this section with illustrations of two additional families of discrete distributions that arise often enough to have names.

Uniform Distributions on Integers

Example 3.1.8

Daily Numbers. A popular state lottery game requires participants to select a three-digit number (leading 0s allowed). Then three balls, each with one digit, are chosen at random from well-mixed bowls. The sample space here consists of all triples (i_1, i_2, i_3) where $i_j \in \{0, \ldots, 9\}$ for $j = 1, 2, 3$. If $s = (i_1, i_2, i_3)$, define $X(s) = 100i_1 + 10i_2 + i_3$. For example, $X(0, 1, 5) = 15$. It is easy to check that $\Pr(X = x) = 0.001$ for each integer $x \in \{0, 1, \ldots, 999\}$. ◀

Definition 3.1.6

Uniform Distribution on Integers. Let $a \le b$ be integers. Suppose that the value of a random variable X is equally likely to be each of the integers $a, \ldots, b$. Then we say that X has the *uniform distribution on the integers* $a, \ldots, b$.

The X in Example 3.1.8 has the uniform distribution on the integers $0, 1, \ldots, 999$. A uniform distribution on a set of k integers has probability $1/k$ on each integer. If $b > a$, there are $b - a + 1$ integers from a to b including a and b. The next result follows immediately from what we have just seen, and it illustrates how the name of the distribution characterizes the distribution.

Theorem 3.1.3

If X has the uniform distribution on the integers $a, \ldots, b$, the p.f. of X is

$$f(x) = \begin{cases} \dfrac{1}{b - a + 1} & \text{for } x = a, \ldots, b, \\ 0 & \text{otherwise.} \end{cases} \qquad \blacksquare$$

The uniform distribution on the integers $a, \ldots, b$ represents the outcome of an experiment that is often described by saying that one of the integers $a, \ldots, b$ is *chosen at random*. In this context, the phrase "at random" means that each of the $b - a + 1$ integers is equally likely to be chosen. In this same sense, it is not possible to choose an integer at random from the set of *all* positive integers, because it is not possible

to assign the same probability to every one of the positive integers and still make the sum of these probabilities equal to 1. In other words, a uniform distribution cannot be assigned to an infinite sequence of possible values, but such a distribution can be assigned to any finite sequence.

Note: Random Variables Can Have the Same Distribution without Being the Same Random Variable. Consider two consecutive daily number draws as in Example 3.1.8. The sample space consists of all 6-tuples $(i_1, \ldots, i_6)$, where the first three coordinates are the numbers drawn on the first day and the last three are the numbers drawn on the second day (all in the order drawn). If $s = (i_1, \ldots, i_6)$, let $X_1(s) = 100i_1 + 10i_2 + i_3$ and let $X_2(s) = 100i_4 + 10i_5 + i_6$. It is easy to see that X_1 and X_2 are different functions of s and are not the same random variable. Indeed, there is only a small probability that they will take the same value. But they have the same distribution because they assume the same values with the same probabilities. If a businessman has 1000 customers numbered $0, \ldots, 999$, and he selects one at random and records the number Y, the distribution of Y will be the same as the distribution of X_1 and of X_2, but Y is not like X_1 or X_2 in any other way.

Binomial Distributions

Example 3.1.9

Defective Parts. Consider again Example 2.2.5 from page 69. In that example, a machine produces a defective item with probability p $(0 < p < 1)$ and produces a nondefective item with probability $1 - p$. We assumed that the events that the different items were defective were mutually independent. Suppose that the experiment consists of examining n of these items. Each outcome of this experiment will consist of a list of which items are defective and which are not, in the order examined. For example, we can let 0 stand for a nondefective item and 1 stand for a defective item. Then each outcome is a string of n digits, each of which is 0 or 1. To be specific, if, say, $n = 6$, then some of the possible outcomes are

$$010010,\ 100100,\ 000011,\ 110000,\ 100001,\ 000000,\ \text{etc.} \tag{3.1.3}$$

We will let X denote the number of these items that are defective. Then the random variable X will have a discrete distribution, and the possible values of X will be $0, 1, 2, \ldots, n$. For example, the first four outcomes listed in Eq. (3.1.3) all have $X(s) = 2$. The last outcome listed has $X(s) = 0$. ◀

Example 3.1.9 is a generalization of Example 2.2.5 with n items inspected rather than just six, and rewritten in the notation of random variables. For $x = 0, 1, \ldots, n$, the probability of obtaining each particular ordered sequence of n items containing exactly x defectives and $n - x$ nondefectives is $p^x(1 - p)^{n-x}$, just as it was in Example 2.2.5. Since there are $\binom{n}{x}$ different ordered sequences of this type, it follows that

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Therefore, the p.f. of X will be as follows:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1 - p)^{n-x} & \text{for } x = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1.4}$$

Definition 3.1.7

Binomial Distribution/Random Variable. The discrete distribution represented by the p.f. in (3.1.4) is called the *binomial distribution with parameters n and p*. A random

variable with this distribution is said to be a *binomial random variable with parameters n and p*.

The reader should be able to verify that the random variable X in Example 3.1.4, the number of heads in a sequence of 10 independent tosses of a fair coin, has the binomial distribution with parameters 10 and 1/2.

Since the name of each binomial distribution is sufficient to construct its p.f., it follows that the name is enough to identify the distribution. The name of each distribution includes the two parameters. The binomial distributions are very important in probability and statistics and will be discussed further in later chapters of this book.

A short table of values of certain binomial distributions is given at the end of this book. It can be found from this table, for example, that if X has the binomial distribution with parameters $n = 10$ and $p = 0.2$, then $\Pr(X = 5) = 0.0264$ and $\Pr(X \geq 5) = 0.0328$.

As another example, suppose that a clinical trial is being run. Suppose that the probability that a patient recovers from her symptoms during the trial is p and that the probability is $1 - p$ that the patient does not recover. Let Y denote the number of patients who recover out of n independent patients in the trial. Then the distribution of Y is also binomial with parameters n and p. Indeed, consider a general experiment that consists of observing n independent repititions (trials) with only two possible results for each trial. For convenience, call the two possible results "success" and "failure." Then the distribution of the number of trials that result in success will be binomial with parameters n and p, where p is the probability of success on each trial.

Note: Names of Distributions. In this section, we gave names to several families of distributions. The name of each distribution includes any numerical parameters that are part of the definition. For example, the random variable X in Example 3.1.4 has the binomial distribution with parameters 10 and 1/2. It is a correct statement to say that X has a binomial distribution or that X has a discrete distribution, but such statements are only partial descriptions of the distribution of X. Such statements are *not* sufficient to name the distribution of X, and hence they are not sufficient as answers to the question "What is the distribution of X?" The same considerations apply to all of the named distributions that we introduce elsewhere in the book. When attempting to specify the distribution of a random variable by giving its name, one must give the full name, including the values of any parameters. Only the full name is sufficient for determining the distribution.

Summary

A random variable is a real-valued function defined on a sample space. The distribution of a random variable X is the collection of all probabilities $\Pr(X \in C)$ for all subsets C of the real numbers such that $\{X \in C\}$ is an event. A random variable X is discrete if there are at most countably many possible values for X. In this case, the distribution of X can be characterized by the probability function (p.f.) of X, namely, $f(x) = \Pr(X = x)$ for x in the set of possible values. Some distributions are so famous that they have names. One collection of such named distributions is the collection of uniform distributions on finite sets of integers. A more famous collection is the collection of binomial distributions whose parameters are n and p, where n is a positive integer and $0 < p < 1$, having p.f. (3.1.4). The binomial distribution with parameters $n = 1$ and p is also called the Bernoulli distribution with parameter p. The names of these distributions also characterize the distributions.

Exercises

1. Suppose that a random variable X has the uniform distribution on the integers $10, \ldots, 20$. Find the probability that X is even.

2. Suppose that a random variable X has a discrete distribution with the following p.f.:

$$f(x) = \begin{cases} cx & \text{for } x = 1, \ldots, 5, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of the constant c.

3. Suppose that two balanced dice are rolled, and let X denote the absolute value of the difference between the two numbers that appear. Determine and sketch the p.f. of X.

4. Suppose that a fair coin is tossed 10 times independently. Determine the p.f. of the number of heads that will be obtained.

5. Suppose that a box contains seven red balls and three blue balls. If five balls are selected at random, without replacement, determine the p.f. of the number of red balls that will be obtained.

6. Suppose that a random variable X has the binomial distribution with parameters $n = 15$ and $p = 0.5$. Find $\Pr(X < 6)$.

7. Suppose that a random variable X has the binomial distribution with parameters $n = 8$ and $p = 0.7$. Find $\Pr(X \geq 5)$ by using the table given at the end of this book. *Hint:* Use the fact that $\Pr(X \geq 5) = \Pr(Y \leq 3)$, where Y has the binomial distribution with parameters $n = 8$ and $p = 0.3$.

8. If 10 percent of the balls in a certain box are red, and if 20 balls are selected from the box at random, with replacement, what is the probability that more than three red balls will be obtained?

9. Suppose that a random variable X has a discrete distribution with the following p.f.:

$$f(x) = \begin{cases} \frac{c}{2^x} & \text{for } x = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of the constant c.

10. A civil engineer is studying a left-turn lane that is long enough to hold seven cars. Let X be the number of cars in the lane at the end of a randomly chosen red light. The engineer believes that the probability that $X = x$ is proportional to $(x + 1)(8 - x)$ for $x = 0, \ldots, 7$ (the possible values of X).

a. Find the p.f. of X.

b. Find the probability that X will be at least 5.

11. Show that there does not exist any number c such that the following function would be a p.f.:

$$f(x) = \begin{cases} \frac{c}{x} & \text{for } x = 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

3.2 Continuous Distributions

Next, we focus on random variables that can assume every value in an interval (bounded or unbounded). If a random variable X has associated with it a function f such that the integral of f over each interval gives the probability that X is in the interval, then we call f the probability density function (p.d.f.) of X and we say that X has a continuous distribution.

The Probability Density Function

Example 3.2.1

Demands for Utilities. In Example 3.1.5, we determined the distribution of the demand for water, X. From Fig. 3.2, we see that the smallest possible value of X is 4 and the largest is 200. For each interval $C = [c_0, c_1] \subset [4, 200]$, Eq. (3.1.2) says that

$$\Pr(c_0 \leq X \leq c_1) = \frac{149(c_1 - c_0)}{29204} = \frac{c_1 - c_0}{196} = \int_{c_0}^{c_1} \frac{1}{196} dx.$$

So, if we define

$$f(x) = \begin{cases} \frac{1}{196} & \text{if } 4 \le x \le 200, \\ 0 & \text{otherwise,} \end{cases} \tag{3.2.1}$$

we have that

$$\Pr(c_0 \le X \le c_1) = \int_{c_0}^{c_1} f(x)dx. \tag{3.2.2}$$

Because we defined $f(x)$ to be 0 for x outside of the interval [4, 200], we see that Eq. (3.2.2) holds for all $c_0 \le c_1$, even if $c_0 = -\infty$ and/or $c_1 = \infty$. ◀

The water demand X in Example 3.2.1 is an example of the following.

Definition 3.2.1 Continuous Distribution/Random Variable. We say that a random variable X has a *continuous distribution* or that X is a *continuous random variable* if there exists a nonnegative function f, defined on the real line, such that for every interval of real numbers (bounded or unbounded), the probability that X takes a value in the interval is the integral of f over the interval.

For example, in the situation described in Definition 3.2.1, for each bounded closed interval $[a, b]$,

$$\Pr(a \le X \le b) = \int_a^b f(x)\,dx. \tag{3.2.3}$$

Similarly, $\Pr(X \ge a) = \int_a^\infty f(x)\,dx$ and $\Pr(X \le b) = \int_{-\infty}^b f(x)\,dx$.

We see that the function f characterizes the distribution of a continuous random variable in much the same way that the probability function characterizes the distribution of a discrete random variable. For this reason, the function f plays an important role, and hence we give it a name.

Definition 3.2.2 Probability Density Function/p.d.f./Support. If X has a continuous distribution, the function f described in Definition 3.2.1 is called the *probability density function* (abbreviated *p.d.f.*) of X. The closure of the set $\{x : f(x) > 0\}$ is called the *support of (the distribution of) X*.

Example 3.2.1 demonstrates that the water demand X has p.d.f. given by (3.2.1).

Every p.d.f. f must satisfy the following two requirements:

$$f(x) \ge 0, \quad \text{for all x,} \tag{3.2.4}$$

and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{3.2.5}$$

A typical p.d.f. is sketched in Fig. 3.4. In that figure, the total area under the curve must be 1, and the value of $\Pr(a \le X \le b)$ is equal to the area of the shaded region.

Note: Continuous Distributions Assign Probability 0 to Individual Values. The integral in Eq. (3.2.3) also equals $\Pr(a < X \le b)$ as well as $\Pr(a < X < b)$ and $\Pr(a \le X < b)$. Hence, it follows from the definition of continuous distributions that, if X has a continuous distribution, $\Pr(X = a) = 0$ for each number a. As we noted on page 20, the fact that $\Pr(X = a) = 0$ does not imply that $X = a$ is impossible. If it did,

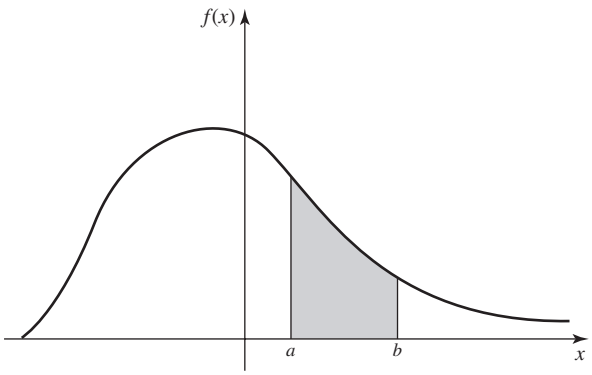

Figure 3.4 An example of a p.d.f.

all values of X would be impossible and X couldn't assume any value. What happens is that the probability in the distribution of X is spread so thinly that we can only see it on sets like nondegenerate intervals. It is much the same as the fact that lines have 0 area in two dimensions, but that does not mean that lines are not there. The two vertical lines indicated under the curve in Fig. 3.4 have 0 area, and this signifies that $\Pr(X = a) = \Pr(X = b) = 0$. However, for each $\epsilon > 0$ and each a such that $f(a) > 0$, $\Pr(a - \epsilon \le X \le a + \epsilon) \approx 2\epsilon f(a) > 0$.

Nonuniqueness of the p.d.f.

If a random variable X has a continuous distribution, then $\Pr(X = x) = 0$ for every individual value x. Because of this property, the values of each p.d.f. can be changed at a finite number of points, or even at certain infinite sequences of points, without changing the value of the integral of the p.d.f. over any subset A. In other words, the values of the p.d.f. of a random variable X can be changed arbitrarily at many points without affecting any probabilities involving X, that is, without affecting the probability distribution of X. At exactly which sets of points we can change a p.d.f. depends on subtle features of the definition of the Riemann integral. We shall not deal with this issue in this text, and we shall only contemplate changes to p.d.f.'s at finitely many points.

To the extent just described, the p.d.f. of a random variable is not unique. In many problems, however, there will be one version of the p.d.f. that is more natural than any other because for this version the p.d.f. will, wherever possible, be continuous on the real line. For example, the p.d.f. sketched in Fig. 3.4 is a continuous function over the entire real line. This p.d.f. could be changed arbitrarily at a few points without affecting the probability distribution that it represents, but these changes would introduce discontinuities into the p.d.f. without introducing any apparent advantages.

Throughout most of this book, we shall adopt the following practice: If a random variable X has a continuous distribution, we shall give only one version of the p.d.f. of X and we shall refer to that version as *the* p.d.f. of X, just as though it had been uniquely determined. It should be remembered, however, that there is some freedom in the selection of the particular version of the p.d.f. that is used to represent each continuous distribution. The most common place where such freedom will arise is in cases like Eq. (3.2.1) where the p.d.f. is required to have discontinuities. Without making the function f any less continuous, we could have defined the p.d.f. in that example so that $f(4) = f(200) = 0$ instead of $f(4) = f(200) = 1/196$. Both of these choices lead to the same calculations of all probabilities associated with X, and they

are both equally valid. Because the support of a continuous distribution is the closure of the set where the p.d.f. is strictly positive, it can be shown that the support is unique. A sensible approach would then be to choose the version of the p.d.f. that was strictly positive on the support whenever possible.

The reader should note that "continuous distribution" is *not* the name of a distribution, just as "discrete distribution" is not the name of a distribution. There are many distributions that are discrete and many that are continuous. Some distributions of each type have names that we either have introduced or will introduce later.

We shall now present several examples of continuous distributions and their p.d.f.'s.

Uniform Distributions on Intervals

Example 3.2.2 Temperature Forecasts. Television weather forecasters announce high and low temperature forecasts as integer numbers of degrees. These forecasts, however, are the results of very sophisticated weather models that provide more precise forecasts that the television personalities round to the nearest integer for simplicity. Suppose that the forecaster announces a high temperature of y. If we wanted to know what temperature X the weather models actually produced, it might be safe to assume that X was equally likely to be any number in the interval from $y - 1/2$ to $y + 1/2$. ◀

The distribution of X in Example 3.2.2 is a special case of the following.

Definition 3.2.3 Uniform Distribution on an Interval. Let a and b be two given real numbers such that $a < b$. Let X be a random variable such that it is known that $a \le X \le b$ and, for every subinterval of $[a, b]$, the probability that X will belong to that subinterval is proportional to the length of that subinterval. We then say that the random variable X has the *uniform distribution on the interval* $[a, b]$.

A random variable X with the uniform distribution on the interval $[a, b]$ represents the outcome of an experiment that is often described by saying that a point is chosen *at random* from the interval $[a, b]$. In this context, the phrase "at random" means that the point is just as likely to be chosen from any particular part of the interval as from any other part of the same length.

Theorem 3.2.1 Uniform Distribution p.d.f. If X has the uniform distribution on an interval $[a, b]$, then the p.d.f. of X is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.6}$$

Proof X must take a value in the interval $[a, b]$. Hence, the p.d.f. $f(x)$ of X must be 0 outside of $[a, b]$. Furthermore, since any particular subinterval of $[a, b]$ having a given length is as likely to contain X as is any other subinterval having the same length, regardless of the location of the particular subinterval in $[a, b]$, it follows that $f(x)$ must be constant throughout $[a, b]$, and that interval is then the support of the distribution. Also,

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{a}^{b} f(x)\,dx = 1. \tag{3.2.7}$$

Therefore, the constant value of $f(x)$ throughout $[a, b]$ must be $1/(b - a)$, and the p.d.f. of X must be (3.2.6). ■

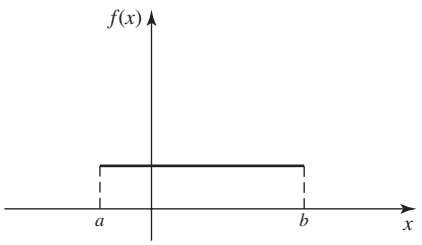

Figure 3.5 The p.d.f. for the uniform distribution on the interval $[a, b]$.

Th p.d.f. (3.2.6) is sketched in Fig. 3.5. As an example, the random variable X (demand for water) in Example 3.2.1 has the uniform distribution on the interval [4, 200].

Note: Density Is Not Probability. The reader should note that the p.d.f. in (3.2.6) can be greater than 1, particularly if $b - a < 1$. Indeed, p.d.f.'s can be unbounded, as we shall see in Example 3.2.6. The p.d.f. of X, $f(x)$, itself does not equal the probability that X is near x. The integral of f over values near x gives the probability that X is near x, and the integral is never greater than 1.

It is seen from Eq. (3.2.6) that the p.d.f. representing a uniform distribution on a given interval is constant over that interval, and the constant value of the p.d.f. is the reciprocal of the length of the interval. It is not possible to define a uniform distribution over an unbounded interval, because the length of such an interval is infinite.

Consider again the uniform distribution on the interval $[a, b]$. Since the probability is 0 that one of the endpoints a or b will be chosen, it is irrelevant whether the distribution is regarded as a uniform distribution on the *closed* interval $a \le x \le b$, or as a uniform distribution on the *open* interval $a < x < b$, or as a uniform distribution on the half-open and half-closed interval $(a, b]$ in which one endpoint is included and the other endpoint is excluded.

For example, if a random variable X has the uniform distribution on the interval $[-1, 4]$, then the p.d.f. of X is

$$f(x) = \begin{cases} 1/5 & \text{for } -1 \le x \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore,

$$\Pr(0 \le X < 2) = \int_0^2 f(x)\,dx = \frac{2}{5}.$$

Notice that we defined the p.d.f. of X to be strictly positive on the closed interval $[-1, 4]$ and 0 outside of this closed interval. It would have been just as sensible to define the p.d.f. to be strictly positive on the open interval $(-1, 4)$ and 0 outside of this open interval. The probability distribution would be the same either way, including the calculation of $\Pr(0 \le X < 2)$ that we just performed. After this, when there are several equally sensible choices for how to define a p.d.f., we will simply choose one of them without making any note of the other choices.

Other Continuous Distributions

Example 3.2.3 Incompletely Specified p.d.f. Suppose that the p.d.f. of a certain random variable X has the following form:

$$f(x) = \begin{cases} cx & \text{for } 0 < x < 4, \\ 0 & \text{otherwise,} \end{cases}$$

where c is a given constant. We shall determine the value of c.

For every p.d.f., it must be true that $\int_{-\infty}^{\infty} f(x) = 1$. Therefore, in this example,

$$\int_0^4 cx\, dx = 8c = 1.$$

Hence, $c = 1/8$. ◀

Note: Calculating Normalizing Constants. The calculation in Example 3.2.3 illustrates an important point that simplifies many statistical results. The p.d.f. of X was specified without explicitly giving the value of the constant c. However, we were able to figure out what was the value of c by using the fact that the integral of a p.d.f. must be 1. It will often happen, especially in Chapter 8 where we find sampling distributions of summaries of observed data, that we can determine the p.d.f. of a random variable except for a constant factor. That constant factor must be the unique value such that the integral of the p.d.f. is 1, even if we cannot calculate it directly.

Example 3.2.4 Calculating Probabilities from a p.d.f. Suppose that the p.d.f. of X is as in Example 3.2.3, namely,

$$f(x) = \begin{cases} \frac{x}{8} & \text{for } 0 < x < 4, \\ 0 & \text{otherwise.} \end{cases}$$

We shall now determine the values of $\Pr(1 \le X \le 2)$ and $\Pr(X > 2)$. Apply Eq. (3.2.3) to get

$$\Pr(1 \le X \le 2) = \int_1^2 \frac{1}{8} x\, dx = \frac{3}{16}$$

and

$$\Pr(X > 2) = \int_2^4 \frac{1}{8} x\, dx = \frac{3}{4}.$$ ◀

Example 3.2.5 Unbounded Random Variables. It is often convenient and useful to represent a continuous distribution by a p.d.f. that is positive over an unbounded interval of the real line. For example, in a practical problem, the voltage X in a certain electrical system might be a random variable with a continuous distribution that can be approximately represented by the p.d.f.

$$f(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \dfrac{1}{(1+x)^2} & \text{for } x > 0. \end{cases} \tag{3.2.8}$$

It can be verified that the properties (3.2.4) and (3.2.5) required of all p.d.f.'s are satisfied by $f(x)$.

Even though the voltage X may actually be bounded in the real situation, the p.d.f. (3.2.8) may provide a good approximation for the distribution of X over its full range of values. For example, suppose that it is known that the maximum possible value of X is 1000, in which case $\Pr(X > 1000) = 0$. When the p.d.f. (3.2.8) is used, we compute $\Pr(X > 1000) = 0.001$. If (3.2.8) adequately represents the variability of X over the interval $(0, 1000)$, then it may be more convenient to use the p.d.f. (3.2.8) than a p.d.f. that is similar to (3.2.8) for $x \le 1000$, except for a new normalizing

constant, and is 0 for $x > 1000$. This can be especially true if we do not know for sure that the maximum voltage is only 1000. ◀

Example 3.2.6

Unbounded p.d.f.'s. Since a value of a p.d.f. is a probability density, rather than a probability, such a value can be larger than 1. In fact, the values of the following p.d.f. are unbounded in the neighborhood of $x = 0$:

$$f(x) = \begin{cases} \frac{2}{3}x^{-1/3} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2.9}$$

It can be verified that even though the p.d.f. (3.2.9) is unbounded, it satisfies the properties (3.2.4) and (3.2.5) required of a p.d.f. ◀

Mixed Distributions

Most distributions that are encountered in practical problems are either discrete or continuous. We shall show, however, that it may sometimes be necessary to consider a distribution that is a mixture of a discrete distribution and a continuous distribution.

Example 3.2.7

Truncated Voltage. Suppose that in the electrical system considered in Example 3.2.5, the voltage X is to be measured by a voltmeter that will record the actual value of X if $X \leq 3$ but will simply record the value 3 if $X > 3$. If we let Y denote the value recorded by the voltmeter, then the distribution of Y can be derived as follows.

First, $\Pr(Y = 3) = \Pr(X \geq 3) = 1/4$. Since the single value $Y = 3$ has probability 1/4, it follows that $\Pr(0 < Y < 3) = 3/4$. Furthermore, since $Y = X$ for $0 < X < 3$, this probability 3/4 for Y is distributed over the interval (0, 3) according to the same p.d.f. (3.2.8) as that of X over the same interval. Thus, the distribution of Y is specified by the combination of a p.d.f. over the interval (0, 3) and a positive probability at the point $Y = 3$. ◀

Summary

A continuous distribution is characterized by its probability density function (p.d.f.). A nonnegative function f is the p.d.f. of the distribution of X if, for every interval $[a, b]$, $\Pr(a \leq X \leq b) = \int_a^b f(x)\,dx$. Continuous random variables satisfy $\Pr(X = x) = 0$ for every value x. If the p.d.f. of a distribution is constant on an interval $[a, b]$ and is 0 off the interval, we say that the distribution is uniform on the interval $[a, b]$.

Exercises

1. Let X be a random variable with the p.d.f. specified in Example 3.2.6. Compute $\Pr(X \leq 8/27)$.

2. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{4}{3}(1 - x^3) & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch this p.d.f. and determine the values of the following probabilities: **a.** $\Pr\left(X < \frac{1}{2}\right)$ **b.** $\Pr\left(\frac{1}{4} < X < \frac{3}{4}\right)$ **c.** $\Pr\left(X > \frac{1}{3}\right)$.

3. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{1}{36}(9 - x^2) & \text{for } -3 \le x \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch this p.d.f. and determine the values of the following probabilities: **a.** $\Pr(X < 0)$ **b.** $\Pr(-1 \le X \le 1)$ **c.** $\Pr(X > 2)$.

4. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} cx^2 & \text{for } 1 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the value of the constant c and sketch the p.d.f.

b. Find the value of $\Pr(X > 3/2)$.

5. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{1}{8}x & \text{for } 0 \le x \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the value of t such that $\Pr(X \le t) = 1/4$.

b. Find the value of t such that $\Pr(X \ge t) = 1/2$.

6. Let X be a random variable for which the p.d.f. is as given in Exercise 5. After the value of X has been observed, let Y be the integer closest to X. Find the p.f. of the random variable Y.

7. Suppose that a random variable X has the uniform distribution on the interval $[-2, 8]$. Find the p.d.f. of X and the value of $\Pr(0 < X < 7)$.

8. Suppose that the p.d.f. of a random variable X is as follows:

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

a. Find the value of the constant c and sketch the p.d.f.

b. Find the value of $\Pr(1 < X < 2)$.

9. Show that there does not exist any number c such that the following function $f(x)$ would be a p.d.f.:

$$f(x) = \begin{cases} \frac{c}{1+x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

10. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{c}{(1-x)^{1/2}} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the value of the constant c and sketch the p.d.f.

b. Find the value of $\Pr(X \le 1/2)$.

11. Show that there does not exist any number c such that the following function $f(x)$ would be a p.d.f.:

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

12. In Example 3.1.3 on page 94, determine the distribution of the random variable Y, the electricity demand. Also, find $\Pr(Y < 50)$.

13. An ice cream seller takes 20 gallons of ice cream in her truck each day. Let X stand for the number of gallons that she sells. The probability is 0.1 that $X = 20$. If she doesn't sell all 20 gallons, the distribution of X follows a continuous distribution with a p.d.f. of the form

$$f(x) = \begin{cases} cx & \text{for } 0 < x < 20, \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant that makes $\Pr(X < 20) = 0.9$. Find the constant c so that $\Pr(X < 20) = 0.9$ as described above.

3.3 The Cumulative Distribution Function

Although a discrete distribution is characterized by its p.f. and a continuous distribution is characterized by its p.d.f., every distribution has a common characterization through its (cumulative) distribution function (c.d.f.). The inverse of the c.d.f. is called the quantile function, and it is useful for indicating where the probability is located in a distribution.

Example 3.3.1

Voltage. Consider again the voltage X from Example 3.2.5. The distribution of X is characterized by the p.d.f. in Eq. (3.2.8). An alternative characterization that is more directly related to probabilities associated with X is obtained from the following function:

$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 0 & \text{for } x \le 0, \\ \displaystyle\int_0^x \frac{dy}{(1+y)^2} & \text{for } x > 0, \end{cases}$$
$$= \begin{cases} 0 & \text{for } x \le 0, \\ 1 - \dfrac{1}{1+x} & \text{for } x > 0. \end{cases} \tag{3.3.1}$$

So, for example, $\Pr(X \le 3) = F(3) = 3/4$. ◀

Definition and Basic Properties

Definition 3.3.1 (Cumulative) Distribution Function. The *distribution function* or *cumulative distribution function* (abbreviated *c.d.f.*) F of a random variable X is the function

$$F(x) = \Pr(X \le x) \quad \text{for } -\infty < x < \infty. \tag{3.3.2}$$

It should be emphasized that the cumulative distribution function is defined as above for every random variable X, regardless of whether the distribution of X is discrete, continuous, or mixed. For the continuous random variable in Example 3.3.1, the c.d.f. was calculated in Eq. (3.3.1). Here is a discrete example:

Example 3.3.2 Bernoulli c.d.f. Let X have the Bernoulli distribution with parameter p defined in Definition 3.1.5. Then $\Pr(X = 0) = 1 - p$ and $\Pr(X = 1) = p$. Let F be the c.d.f. of X. It is easy to see that $F(x) = 0$ for $x < 0$ because $X \ge 0$ for sure. Similarly, $F(x) = 1$ for $x \ge 1$ because $X \le 1$ for sure. For $0 \le x < 1$, $\Pr(X \le x) = \Pr(X = 0) = 1 - p$ because 0 is the only possible value of X that is in the interval $(-\infty, x]$. In summary,

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 - p & \text{for } 0 \le x < 1, \\ 1 & \text{for } x \ge 1. \end{cases}$$ ◀

We shall soon see (Theorem 3.3.2) that the c.d.f. allows calculation of all interval probabilities; hence, it characterizes the distribution of a random variable. It follows from Eq. (3.3.2) that the c.d.f. of each random variable X is a function F defined on the real line. The value of F at every point x must be a number $F(x)$ in the interval $[0, 1]$ because $F(x)$ is the probability of the event $\{X \le x\}$. Furthermore, it follows from Eq. (3.3.2) that the c.d.f. of every random variable X must have the following three properties.

Property 3.3.1 Nondecreasing. *The function $F(x)$ is nondecreasing as x increases; that is, if $x_1 < x_2$, then $F(x_1) \le F(x_2)$.*

Proof If $x_1 < x_2$, then the event $\{X \le x_1\}$ is a subset of the event $\{X \le x_2\}$. Hence, $\Pr\{X \le x_1\} \le \Pr\{X \le x_2\}$ according to Theorem 1.5.4. ■

An example of a c.d.f. is sketched in Fig. 3.6. It is shown in that figure that $0 \le F(x) \le 1$ over the entire real line. Also, $F(x)$ is always nondecreasing as x increases, although $F(x)$ is constant over the interval $x_1 \le x \le x_2$ and for $x \ge x_4$.

Property 3.3.2 Limits at $\pm\infty$. $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.

Proof As in the proof of Property 3.3.1, note that $\{X \le x_1\} \subset \{X \le x_2\}$ whenever $x_1 < x_2$. The fact that $\Pr(X \le x)$ approaches 0 as $x \to -\infty$ now follows from Exercise 13 in

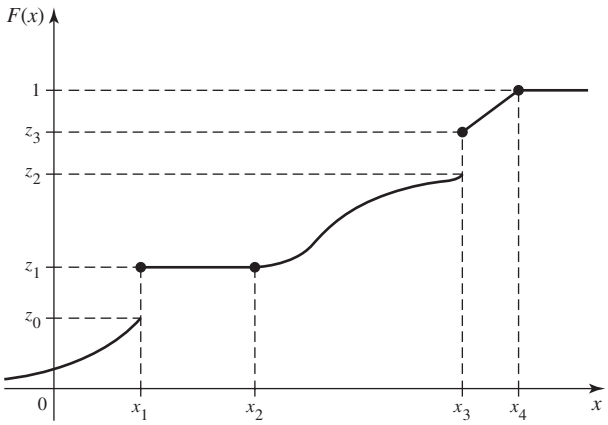

Figure 3.6 An example of a c.d.f.

Section 1.10. Similarly, the fact that $\Pr(X \le x)$ approaches 1 as $x \to \infty$ follows from Exercise 12 in Sec. 1.10. ■

The limiting values specified in Property 3.3.2 are indicated in Fig. 3.6. In this figure, the value of $F(x)$ actually becomes 1 at $x = x_4$ and then remains 1 for $x > x_4$. Hence, it may be concluded that $\Pr(X \le x_4) = 1$ and $\Pr(X > x_4) = 0$. On the other hand, according to the sketch in Fig. 3.6, the value of $F(x)$ approaches 0 as $x \to -\infty$, but does not actually become 0 at any finite point x. Therefore, for every finite value of x, no matter how small, $\Pr(X \le x) > 0$.

A c.d.f. need not be continuous. In fact, the value of $F(x)$ may jump at any finite or countable number of points. In Fig. 3.6, for instance, such jumps or points of discontinuity occur where $x = x_1$ and $x = x_3$. For each fixed value x, we shall let $F(x^-)$ denote the limit of the values of $F(y)$ as y approaches x from the left, that is, as y approaches x through values smaller than x. In symbols,

$$F(x^-) = \lim_{\substack{y \to x \\ y < x}} F(y).$$

Similarly, we shall define $F(x^+)$ as the limit of the values of $F(y)$ as y approaches x from the right. Thus,

$$F(x^+) = \lim_{\substack{y \to x \\ y > x}} F(y).$$

If the c.d.f. is continuous at a given point x, then $F(x^-) = F(x^+) = F(x)$ at that point.

Property 3.3.3 Continuity from the Right. *A c.d.f. is always continuous from the right; that is,* $F(x) = F(x^+)$ *at every point* x.

Proof Let $y_1 > y_2 > \cdots$ be a sequence of numbers that are decreasing such that $\lim_{n\to\infty} y_n = x$. Then the event $\{X \le x\}$ is the intersection of all the events $\{X \le y_n\}$ for $n = 1, 2, \ldots$. Hence, by Exercise 13 of Sec. 1.10,

$$F(x) = \Pr(X \le x) = \lim_{n\to\infty} \Pr(X \le y_n) = F(x^+).$$ ■

It follows from Property 3.3.3 that at every point x at which a jump occurs,

$$F(x^+) = F(x) \text{ and } F(x^-) < F(x).$$

In Fig. 3.6 this property is illustrated by the fact that, at the points of discontinuity $x = x_1$ and $x = x_3$, the value of $F(x_1)$ is taken as z_1 and the value of $F(x_3)$ is taken as z_3.

Determining Probabilities from the Distribution Function

Example 3.3.3 Voltage. In Example 3.3.1, suppose that we want to know the probability that X lies in the interval $[2, 4]$. That is, we want $\Pr(2 \le X \le 4)$. The c.d.f. allows us to compute $\Pr(X \le 4)$ and $\Pr(X \le 2)$. These are related to the probability that we want as follows: Let $A = \{2 < X \le 4\}$, $B = \{X \le 2\}$, and $C = \{X \le 4\}$. Because X has a continuous distribution, $\Pr(A)$ is the same as the probability that we desire. We see that $A \cup B = C$, and it is clear that A and B are disjoint. Hence, $\Pr(A) + \Pr(B) = \Pr(C)$. It follows that

$$\Pr(A) = \Pr(C) - \Pr(B) = F(4) - F(2) = \frac{4}{5} - \frac{3}{4} = \frac{1}{20}. \quad \blacktriangleleft$$

The type of reasoning used in Example 3.3.3 can be extended to find the probability that an arbitrary random variable X will lie in any specified interval of the real line from the c.d.f. We shall derive this probability for four different types of intervals.

Theorem 3.3.1 For every value x,

$$\Pr(X > x) = 1 - F(x). \tag{3.3.3}$$

Proof The events $\{X > x\}$ and $\{X \le x\}$ are disjoint, and their union is the whole sample space S whose probability is 1. Hence, $\Pr(X > x) + \Pr(X \le x) = 1$. Now, Eq. (3.3.3) follows from Eq. (3.3.2). ■

Theorem 3.3.2 For all values x_1 and x_2 such that $x_1 < x_2$,

$$\Pr(x_1 < X \le x_2) = F(x_2) - F(x_1). \tag{3.3.4}$$

Proof Let $A = \{x_1 < X \le x_2\}$, $B = \{X \le x_1\}$, and $C = \{X \le x_2\}$. As in Example 3.3.3, A and B are disjoint, and their union is C, so

$$\Pr(x_1 < X \le x_2) + \Pr(X \le x_1) = \Pr(X \le x_2).$$

Subtracting $\Pr(X \le x_1)$ from both sides of this equation and applying Eq. (3.3.2) yields Eq. (3.3.4). ■

For example, if the c.d.f. of X is as sketched in Fig. 3.6, then it follows from Theorems 3.3.1 and 3.3.2 that $\Pr(X > x_2) = 1 - z_1$ and $\Pr(x_2 < X \le x_3) = z_3 - z_1$. Also, since $F(x)$ is constant over the interval $x_1 \le x \le x_2$, then $\Pr(x_1 < X \le x_2) = 0$.

It is important to distinguish carefully between the strict inequalities and the weak inequalities that appear in all of the preceding relations and also in the next theorem. If there is a jump in $F(x)$ at a given value x, then the values of $\Pr(X \le x)$ and $\Pr(X < x)$ will be different.

Theorem 3.3.3 For each value x,

$$\Pr(X < x) = F(x^-). \tag{3.3.5}$$

Proof Let $y_1 < y_2 < \cdots$ be an increasing sequence of numbers such that $\lim_{n\to\infty} y_n = x$. Then it can be shown that

$$\{X < x\} = \bigcup_{n=1}^{\infty} \{X \le y_n\}.$$

Therefore, it follows from Exercise 12 of Sec. 1.10 that

$$\begin{aligned} \Pr(X < x) &= \lim_{n\to\infty} \Pr(X \le y_n) \\ &= \lim_{n\to\infty} F(y_n) = F(x^-). \end{aligned}$$

■

For example, for the c.d.f. sketched in Fig. 3.6, $\Pr(X < x_3) = z_2$ and $\Pr(X < x_4) = 1$.

Finally, we shall show that for every value x, $\Pr(X = x)$ is equal to the amount of the jump that occurs in F at the point x. If F is continuous at the point x, that is, if there is no jump in F at x, then $\Pr(X = x) = 0$.

Theorem 3.3.4 For every value x,

$$\Pr(X = x) = F(x) - F(x^-). \tag{3.3.6}$$

Proof It is always true that $\Pr(X = x) = \Pr(X \le x) - \Pr(X < x)$. The relation (3.3.6) follows from the fact that $\Pr(X \le x) = F(x)$ at every point and from Theorem 3.3.3. ■

In Fig. 3.6, for example, $\Pr(X = x_1) = z_1 - z_0$, $\Pr(X = x_3) = z_3 - z_2$, and the probability of every other individual value of X is 0.

The c.d.f. of a Discrete Distribution

From the definition and properties of a c.d.f. $F(x)$, it follows that if $a < b$ and if $\Pr(a < X < b) = 0$, then $F(x)$ will be constant and horizontal over the interval $a < x < b$. Furthermore, as we have just seen, at every point x such that $\Pr(X = x) > 0$, the c.d.f. will jump by the amount $\Pr(X = x)$.

Suppose that X has a discrete distribution with the p.f. $f(x)$. Together, the properties of a c.d.f. imply that $F(x)$ must have the following form: $F(x)$ will have a jump of magnitude $f(x_i)$ at each possible value x_i of X, and $F(x)$ will be constant between every pair of successive jumps. The distribution of a discrete random variable X can be represented equally well by either the p.f. or the c.d.f. of X.

The c.d.f. of a Continuous Distribution

Theorem 3.3.5 Let X have a continuous distribution, and let $f(x)$ and $F(x)$ denote its p.d.f. and the c.d.f., respectively. Then F is continuous at every x,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \tag{3.3.7}$$

and

$$\frac{dF(x)}{dx} = f(x), \tag{3.3.8}$$

at all x such that f is continuous.

Proof Since the probability of each individual point x is 0, the c.d.f. $F(x)$ will have no jumps. Hence, $F(x)$ will be a continuous function over the entire real line.

By definition, $F(x) = \Pr(X \le x)$. Since f is the p.d.f. of X, we have from the definition of p.d.f. that $\Pr(X \le x)$ is the right-hand side of Eq. (3.3.7).

It follows from Eq. (3.3.7) and the relation between integrals and derivatives (the fundamental theorem of calculus) that, for every x at which f is continuous, Eq. (3.3.8) holds. ■

Thus, the c.d.f. of a continuous random variable X can be obtained from the p.d.f. and vice versa. Eq. (3.3.7) is how we found the c.d.f. in Example 3.3.1. Notice that the derivative of the F in Example 3.3.1 is

$$F'(x) = \begin{cases} 0 & \text{for } x < 0, \\ \dfrac{1}{(1+x)^2} & \text{for } x > 0, \end{cases}$$

and F' does not exist at $x = 0$. This verifies Eq (3.3.8) for Example 3.3.1. Here, we have used the popular shorthand notation $F'(x)$ for the derivative of F at the point x.

Example 3.3.4 Calculating a p.d.f. from a c.d.f. Let the c.d.f. of a random variable be

$$F(x) = \begin{cases} 0 & \text{for } x < 0, \\ x^{2/3} & \text{for } 0 \le x \le 1, \\ 1 & \text{for } x > 1. \end{cases}$$

This function clearly satisfies the three properties required of every c.d.f., as given earlier in this section. Furthermore, since this c.d.f. is continuous over the entire real line and is differentiable at every point except $x = 0$ and $x = 1$, the distribution of X is continuous. Therefore, the p.d.f. of X can be found at every point other than $x = 0$ and $x = 1$ by the relation (3.3.8). The value of $f(x)$ at the points $x = 0$ and $x = 1$ can be assigned arbitrarily. When the derivative $F'(x)$ is calculated, it is found that $f(x)$ is as given by Eq. (3.2.9) in Example 3.2.6. Conversely, if the p.d.f. of X is given by Eq. (3.2.9), then by using Eq. (3.3.7) it is found that $F(x)$ is as given in this example. ◀

The Quantile Function

Example 3.3.5 Fair Bets. Suppose that X is the amount of rain that will fall tomorrow, and X has c.d.f. F. Suppose that we want to place an even-money bet on X as follows: If $X \le x_0$, we win one dollar and if $X > x_0$ we lose one dollar. In order to make this bet fair, we need $\Pr(X \le x_0) = \Pr(X > x_0) = 1/2$. We could search through all of the real numbers x trying to find one such that $F(x) = 1/2$, and then we would let x_0 equal the value we found. If F is a one-to-one function, then F has an inverse F^{-1} and $x_0 = F^{-1}(1/2)$. ◀

The value x_0 that we seek in Example 3.3.5 is called the 0.5 *quantile* of X or the 50th *percentile* of X because 50% of the distribution of X is at or below x_0.

Definition 3.3.2 Quantiles/Percentiles. Let X be a random variable with c.d.f. F. For each p strictly between 0 and 1, define $F^{-1}(p)$ to be the smallest value x such that $F(x) \ge p$. Then $F^{-1}(p)$ is called the p *quantile* of X or the $100p$ *percentile* of X. The function F^{-1} defined here on the open interval (0, 1) is called the *quantile function* of X.

Example 3.3.6

Standardized Test Scores. Many universities in the United States rely on standardized test scores as part of their admissions process. Thousands of people take these tests each time that they are offered. Each examinee's score is compared to the collection of scores of all examinees to see where it fits in the overall ranking. For example, if 83% of all test scores are at or below your score, your test report will say that you scored at the 83rd percentile. ◀

The notation $F^{-1}(p)$ in Definition 3.3.2 deserves some justification. Suppose first that the c.d.f. F of X is continuous and one-to-one over the whole set of possible values of X. Then the inverse F^{-1} of F exists, and for each $0 < p < 1$, there is one and only one x such that $F(x) = p$. That x is $F^{-1}(p)$. Definition 3.3.2 extends the concept of inverse function to nondecreasing functions (such as c.d.f.'s) that may be neither one-to-one nor continuous.

Quantiles of Continuous Distributions When the c.d.f. of a random variable X is continuous and one-to-one over the whole set of possible values of X, the inverse F^{-1} of F exists and equals the quantile function of X.

Example 3.3.7

Value at Risk. The manager of an investment portfolio is interested in how much money the portfolio might lose over a fixed time horizon. Let X be the change in value of the given portfolio over a period of one month. Suppose that X has the p.d.f. in Fig. 3.7. The manager computes a quantity known in the world of risk management as *Value at Risk* (denoted by VaR). To be specific, let $Y = -X$ stand for the loss incurred by the portfolio over the one month. The manager wants to have a level of confidence about how large Y might be. In this example, the manager specifies a probability level, such as 0.99 and then finds y_0, the 0.99 quantile of Y. The manager is now 99% sure that $Y \le y_0$, and y_0 is called the VaR. If X has a continuous distribution, then it is easy to see that y_0 is closely related to the 0.01 quantile of the distribution of X. The 0.01 quantile x_0 has the property that $\Pr(X < x_0) = 0.01$. But $\Pr(X < x_0) = \Pr(Y > -x_0) = 1 - \Pr(Y \le -x_0)$. Hence, $-x_0$ is a 0.99 quantile of Y. For the p.d.f. in Fig. 3.7, we see that $x_0 = -4.14$, as the shaded region indicates. Then $y_0 = 4.14$ is VaR for one month at probability level 0.99. ◀

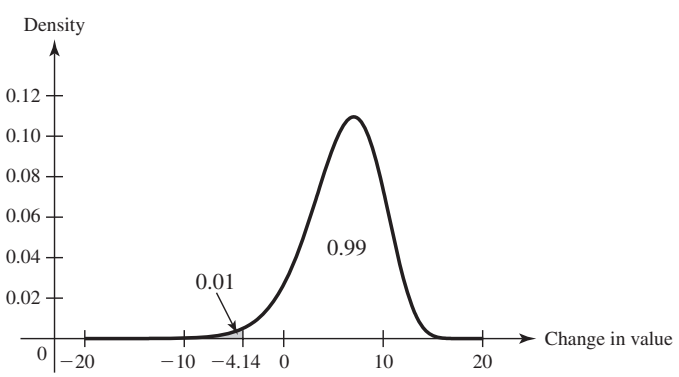

Figure 3.7 The p.d.f. of the change in value of a portfolio with lower 1% indicated.

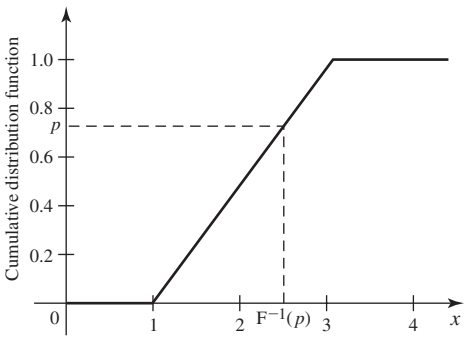

Figure 3.8 The c.d.f. of a uniform distribution indicating how to solve for a quantile.

Example 3.3.8 Uniform Distribution on an Interval. Let X have the uniform distribution on the interval $[a, b]$. The c.d.f. of X is

$$F(x) = \Pr(X \le x) = \begin{cases} 0 & \text{if } x \le a, \\ \displaystyle\int_a^x \frac{1}{b-a} du & \text{if } a < x \le b, \\ 1 & \text{if } x > b. \end{cases}$$

The integral above equals $(x-a)/(b-a)$. So, $F(x) = (x-a)/(b-a)$ for all $a < x < b$, which is a strictly increasing function over the entire interval of possible values of X. The inverse of this function is the quantile function of X, which we obtain by setting $F(x)$ equal to p and solving for x:

$$\begin{aligned} \frac{x-a}{b-a} &= p, \\ x - a &= p(b-a), \\ x &= a + p(b-a) = pb + (1-p)a. \end{aligned}$$

Figure 3.8 illustrates how the calculation of a quantile relates to the c.d.f.

The quantile function of X is $F^{-1}(p) = pb + (1-p)a$ for $0 < p < 1$. In particular, $F^{-1}(1/2) = (b+a)/2$. ◀

Note: Quantiles, Like c.d.f.'s, Depend on the Distribution Only. Any two random variables with the same distribution have the same quantile function. When we refer to a quantile of X, we mean a quantile of the distribution of X.

Quantiles of Discrete Distributions It is convenient to be able to calculate quantiles for discrete distributions as well. The quantile function of Definition 3.3.2 exists for all distributions whether discrete, continuous, or otherwise. For example, in Fig. 3.6, let $z_0 \le p \le z_1$. Then the smallest x such that $F(x) \ge p$ is x_1. For every value of $x < x_1$, we have $F(x) < z_0 \le p$ and $F(x_1) = z_1$. Notice that $F(x) = z_1$ for all x between x_1 and x_2, but since x_1 is the smallest of all those numbers, x_1 is the p quantile. Because distribution functions are continuous from the right, the smallest x such that $F(x) \ge p$ exists for all $0 < p < 1$. For $p = 1$, there is no guarantee that such an x will exist. For example, in Fig. 3.6, $F(x_4) = 1$, but in Example 3.3.1, $F(x) < 1$ for all x. For $p = 0$, there is never a smallest x such that $F(x) = 0$ because $\lim_{x \to -\infty} F(x) = 0$. That is, if $F(x_0) = 0$, then $F(x) = 0$ for all $x < x_0$. For these reasons, we never talk about the 0 or 1 quantiles.

Table 3.1 Quantile function for Example 3.3.9

p	$F^{-1}(p)$
(0, 0.1681]	0
(0.1681, 0.5283]	1
(0.5283, 0.8370]	2
(0.8370, 0.9693]	3
(0.9693, 0.9977]	4
(0.9977, 1)	5

Example 3.3.9

Quantiles of a Binomial Distribution. Let X have the binomial distribution with parameters 5 and 0.3. The binomial table in the back of the book has the p.f. f of X, which we reproduce here together with the c.d.f. F:

x	0	1	2	3	4	5
$f(x)$	0.1681	0.3602	0.3087	0.1323	0.0284	0.0024
$F(x)$	0.1681	0.5283	0.8370	0.9693	0.9977	1

(A little rounding error occurred in the p.f.) So, for example, the 0.5 quantile of this distribution is 1, which is also the 0.25 quantile and the 0.20 quantile. The entire quantile function is in Table 3.1. So, the 90th percentile is 3, which is also the 95th percentile, etc. ◀

Certain quantiles have special names.

Definition 3.3.3

Median/Quartiles. The 1/2 quantile or the 50th percentile of a distribution is called its *median*. The 1/4 quantile or 25th percentile is the *lower quartile*. The 3/4 quantile or 75th percentile is called the *upper quartile*.

Note: The Median Is Special. The median of a distribution is one of several special features that people like to use when sumarizing the distribution of a random variable. We shall discuss summaries of distributions in more detail in Chapter 4. Because the median is such a popular summary, we need to note that there are several different but similar "definitions" of median. Recall that the 1/2 quantile is the *smallest* number x such that $F(x) \geq 1/2$. For some distributions, usually discrete distributions, there will be an interval of numbers $[x_1, x_2)$ such that for all $x \in [x_1, x_2)$, $F(x) = 1/2$. In such cases, it is common to refer to all such x (including x_2) as medians of the distribution. (See Definition 4.5.1.) Another popular convention is to call $(x_1 + x_2)/2$ the median. This last is probably the most common convention. The readers should be aware that, whenever they encounter a median, it might be any one of the things that we just discussed. Fortunately, they all mean nearly the same thing, namely that the number divides the distribution in half as closely as is possible.

Example 3.3.10 Uniform Distribution on Integers. Let X have the uniform distribution on the integers 1, 2, 3, 4. (See Definition 3.1.6.) The c.d.f. of X is

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1/4 & \text{if } 1 \le x < 2, \\ 1/2 & \text{if } 2 \le x < 3, \\ 3/4 & \text{if } 3 \le x < 4, \\ 1 & \text{if } x \ge 4. \end{cases}$$

The 1/2 quantile is 2, but every number in the interval [2, 3] might be called a median. The most popular choice would be 2.5. ◀

One advantage to describing a distribution by the quantile function rather than by the c.d.f. is that quantile functions are easier to display in tabular form for multiple distributions. The reason is that the domain of the quantile function is always the interval (0, 1) no matter what the possible values of X are. Quantiles are also useful for summarizing distributions in terms of where the probability is. For example, if one wishes to say where the middle half of a distribution is, one can say that it lies between the 0.25 quantile and the 0.75 quantile. In Sec. 8.5, we shall see how to use quantiles to help provide estimates of unknown quantities after observing data.

In Exercise 19, you can show how to recover the c.d.f. from the quantile function. Hence, the quantile function is an alternative way to characterize a distribution.

Summary

The c.d.f. F of a random variable X is $F(x) = \Pr(X \le x)$ for all real x. This function is continuous from the right. If we let $F(x^-)$ equal the limit of $F(y)$ as y approaches x from below, then $F(x) - F(x^-) = \Pr(X = x)$. A continuous distribution has a continuous c.d.f. and $F'(x) = f(x)$, the p.d.f. of the distribution, for all x at which F is differentiable. A discrete distribution has a c.d.f. that is constant between the possible values and jumps by $f(x)$ at each possible value x. The quantile function $F^{-1}(p)$ is equal to the smallest x such that $F(x) \ge p$ for $0 < p < 1$.

Exercises

1. Suppose that a random variable X has the Bernoulli distribution with parameter $p = 0.7$. (See Definition 3.1.5.) Sketch the c.d.f. of X.

2. Suppose that a random variable X can take only the values -2, 0, 1, and 4, and that the probabilities of these values are as follows: $\Pr(X = -2) = 0.4$, $\Pr(X = 0) = 0.1$, $\Pr(X = 1) = 0.3$, and $\Pr(X = 4) = 0.2$. Sketch the c.d.f. of X.

3. Suppose that a coin is tossed repeatedly until a head is obtained for the first time, and let X denote the number of tosses that are required. Sketch the c.d.f. of X.

4. Suppose that the c.d.f. F of a random variable X is as sketched in Fig. 3.9. Find each of the following probabilities:

a. $\Pr(X = -1)$ **b.** $\Pr(X < 0)$
c. $\Pr(X \le 0)$ **d.** $\Pr(X = 1)$
e. $\Pr(0 < X \le 3)$ **f.** $\Pr(0 < X < 3)$
g. $\Pr(0 \le X \le 3)$ **h.** $\Pr(1 < X \le 2)$
i. $\Pr(1 \le X \le 2)$ **j.** $\Pr(X > 5)$
k. $\Pr(X \ge 5)$ **l.** $\Pr(3 \le X \le 4)$

5. Suppose that the c.d.f. of a random variable X is as follows:

$$F(x) = \begin{cases} 0 & \text{for } x \le 0, \\ \frac{1}{9}x^2 & \text{for } 0 < x \le 3, \\ 1 & \text{for } x > 3. \end{cases}$$

Find and sketch the p.d.f. of X.

6. Suppose that the c.d.f. of a random variable X is as follows:

$$F(x) = \begin{cases} e^{x-3} & \text{for } x \le 3, \\ 1 & \text{for } x > 3. \end{cases}$$

Find and sketch the p.d.f. of X.

7. Suppose, as in Exercise 7 of Sec. 3.2, that a random variable X has the uniform distribution on the interval $[-2, 8]$. Find and sketch the c.d.f. of X.

8. Suppose that a point in the xy-plane is chosen at random from the interior of a circle for which the equation is $x^2 + y^2 = 1$; and suppose that the probability that the point will belong to each region inside the circle is proportional to the area of that region. Let Z denote a random variable representing the distance from the center of the circle to the point. Find and sketch the c.d.f. of Z.

9. Suppose that X has the uniform distribution on the interval $[0, 5]$ and that the random variable Y is defined by $Y = 0$ if $X \le 1$, $Y = 5$ if $X \ge 3$, and $Y = X$ otherwise. Sketch the c.d.f. of Y.

10. For the c.d.f. in Example 3.3.4, find the quantile function.

11. For the c.d.f. in Exercise 5, find the quantile function.

12. For the c.d.f. in Exercise 6, find the quantile function.

13. Suppose that a broker believes that the change in value X of a particular investment over the next two months has the uniform distribution on the interval $[-12, 24]$. Find the value at risk VaR for two months at probability level 0.95.

14. Find the quartiles and the median of the binomial distribution with parameters $n = 10$ and $p = 0.2$.

15. Suppose that X has the p.d.f.

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

Find and sketch the c.d.f. or X.

16. Find the quantile function for the distribution in Example 3.3.1.

17. Prove that the quantile function F^{-1} of a general random variable X has the following three properties that are analogous to properties of the c.d.f.:

a. F^{-1} is a nondecreasing function of p for $0 < p < 1$.

b. Let $x_0 = \lim_{\substack{p \to 0 \\ p > 0}} F^{-1}(p)$ and $x_1 = \lim_{\substack{p \to 1 \\ p < 1}} F^{-1}(p)$. Then x_0 equals the greatest lower bound on the set of numbers c such that $\Pr(X \le c) > 0$, and x_1 equals the least upper bound on the set of numbers d such that $\Pr(X \ge d) > 0$.

c. F^{-1} is continuous from the left; that is $F^{-1}(p) = F^{-1}(p^-)$ for all $0 < p < 1$.

18. Let X be a random variable with quantile function F^{-1}. Assume the following three conditions: (i) $F^{-1}(p) = c$ for all p in the interval (p_0, p_1), (ii) either $p_0 = 0$ or $F^{-1}(p_0) < c$, and (iii) either $p_1 = 1$ or $F^{-1}(p) > c$ for $p > p_1$. Prove that $\Pr(X = c) = p_1 - p_0$.

19. Let X be a random variable with c.d.f. F and quantile function F^{-1}. Let x_0 and x_1 be as defined in Exercise 17. (Note that $x_0 = -\infty$ and/or $x_1 = \infty$ are possible.) Prove that for all x in the open interval (x_0, x_1), $F(x)$ is the largest p such that $F^{-1}(p) \le x$.

20. In Exercise 13 of Sec. 3.2, draw a sketch of the c.d.f. F of X and find $F(10)$.

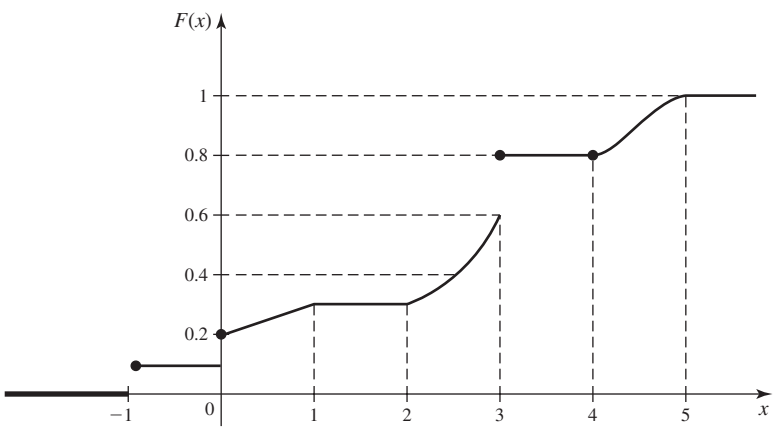

Figure 3.9 The c.d.f. for Exercise 4.

3.4 Bivariate Distributions

We generalize the concept of distribution of a random variable to the joint distribution of two random variables. In doing so, we introduce the joint p.f. for two discrete random variables, the joint p.d.f. for two continuous random variables, and the joint c.d.f. for any two random variables. We also introduce a joint hybrid of p.f. and p.d.f. for the case of one discrete random variable and one continuous random variable.

Example 3.4.1 Demands for Utilities. In Example 3.1.5, we found the distribution of the random variable X that represented the demand for water. But there is another random variable, Y, the demand for electricity, that is also of interest. When discussing two random variables at once, it is often convenient to put them together into an ordered pair, (X, Y). As early as Example 1.5.4 on page 19, we actually calculated some probabilities associated with the pair (X, Y). In that example, we defined two events A and B that we now can express as $A = \{X \geq 115\}$ and $B = \{Y \geq 110\}$. In Example 1.5.4, we computed $\Pr(A \cap B)$ and $\Pr(A \cup B)$. We can express $A \cap B$ and $A \cup B$ as events involving the pair (X, Y). For example, define the set of ordered pairs $C = \{(x, y) : x \geq 115 \text{ and } y \geq 110\}$ so that that the event $\{(X, Y) \in C)\} = A \cap B$. That is, the event that the pair of random variables lies in the set C is the same as the intersection of the two events A and B. In Example 1.5.4, we computed $\Pr(A \cap B) = 0.1198$. So, we can now assert that $\Pr((X, Y) \in C) = 0.1198$. ◀

Definition 3.4.1 Joint/Bivariate Distribution. Let X and Y be random variables. The *joint distribution* or *bivariate distribution* of X and Y is the collection of all probabilities of the form $\Pr[(X, Y) \in C]$ for all sets C of pairs of real numbers such that $\{(X, Y) \in C\}$ is an event.

It is a straightforward consequence of the definition of the joint distribution of X and Y that this joint distribution is itself a probability measure on the set of ordered pairs of real numbers. The set $\{(X, Y) \in C\}$ will be an event for every set C of pairs of real numbers that most readers will be able to imagine.

In this section and the next two sections, we shall discuss convenient ways to characterize and do computations with bivariate distributions. In Sec. 3.7, these considerations will be extended to the joint distribution of an arbitrary finite number of random variables.

Discrete Joint Distributions

Example 3.4.2 Theater Patrons. Suppose that a sample of 10 people is selected at random from a theater with 200 patrons. One random variable of interest might be the number X of people in the sample who are over 60 years of age, and another random variable might be the number Y of people in the sample who live more than 25 miles from the theater. For each ordered pair (x, y) with $x = 0, \ldots, 10$ and $y = 0, \ldots, 10$, we might wish to compute $\Pr((X, Y) = (x, y))$, the probability that there are x people in the sample who are over 60 years of age and there are y people in the sample who live more than 25 miles away. ◀

Definition 3.4.2 Discrete Joint Distribution. Let X and Y be random variables, and consider the ordered pair (X, Y). If there are only finitely or at most countably many different possible values (x, y) for the pair (X, Y), then we say that X and Y have a *discrete joint distribution*.

The two random variables in Example 3.4.2 have a discrete joint distribution.

Theorem 3.4.1 Suppose that two random variables X and Y each have a discrete distribution. Then X and Y have a discrete joint distribution.

Proof If both X and Y have only finitely many possible values, then there will be only a finite number of different possible values (x, y) for the pair (X, Y). On the other hand, if either X or Y or both can take a countably infinite number of possible values, then there will also be a countably infinite number of possible values for the pair (X, Y). In all of these cases, the pair (X, Y) has a discrete joint distribution. ■

When we define continuous joint distribution shortly, we shall see that the obvious analog of Theorem 3.4.1 is not true.

Definition 3.4.3 Joint Probability Function, p.f. The *joint probability function*, or the *joint p.f.*, of X and Y is defined as the function f such that for every point (x, y) in the xy-plane,

$$f(x, y) = \Pr(X = x \text{ and } Y = y).$$

The following result is easy to prove because there are at most countably many pairs (x, y) that must account for all of the probability a discrete joint distribution.

Theorem 3.4.2 Let X and Y have a discrete joint distribution. If (x, y) is not one of the possible values of the pair (X, Y), then $f(x, y) = 0$. Also,

$$\sum_{\text{All } (x,y)} f(x, y) = 1.$$

Finally, for each set C of ordered pairs,

$$\Pr[(X, Y) \in C] = \sum_{(x,y)\in C} f(x, y).$$ ■

Example 3.4.3 Specifying a Discrete Joint Distribution by a Table of Probabilities. In a certain suburban area, each household reported the number of cars and the number of television sets that they owned. Let X stand for the number of cars owned by a randomly selected household in this area. Let Y stand for the number of television sets owned by that same randomly selected household. In this case, X takes only the values 1, 2, and 3; Y takes only the values 1, 2, 3, and 4; and the joint p.f. f of X and Y is as specified in Table 3.2.

Table 3.2 Joint p.f. $f(x, y)$ for Example 3.4.3

	y			
x	1	2	3	4
1	0.1	0	0.1	0
2	0.3	0	0.1	0.2
3	0	0.2	0	0

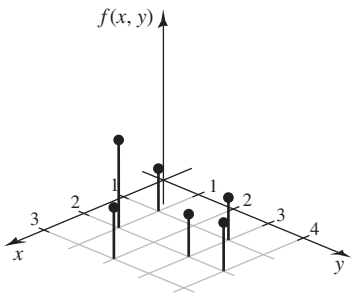

Figure 3.10 The joint p.f. of X and Y in Example 3.4.3.

This joint p.f. is sketched in Fig. 3.10. We shall determine the probability that the randomly selected household owns at least two of both cars and televisions. In symbols, this is $\Pr(X \geq 2 \text{ and } Y \geq 2)$.

By summing $f(x, y)$ over all values of $x \geq 2$ and $y \geq 2$, we obtain the value

$$\begin{aligned}\Pr(X \geq 2 \text{ and } Y \geq 2) &= f(2, 2) + f(2, 3) + f(2, 4) + f(3, 2) \\ &\quad + f(3, 3) + f(3, 4) \\ &= 0.5.\end{aligned}$$

Next, we shall determine the probability that the randomly selected household owns exactly one car, namely $\Pr(X = 1)$. By summing the probabilities in the first row of the table, we obtain the value

$$\Pr(X = 1) = \sum_{y=1}^{4} f(1, y) = 0.2. \quad \blacktriangleleft$$

Continuous Joint Distributions

Example 3.4.4 Demands for Utilities. Consider again the joint distribution of X and Y in Example 3.4.1. When we first calculated probabilities for these two random variables back in Example 1.5.4 on page 19 (even before we named them or called them random variables), we assumed that the probability of each subset of the sample space was proportional to the area of the subset. Since the area of the sample space is 29,204, the probability that the pair (X, Y) lies in a region C is the area of C divided by 29,204. We can also write this relation as

$$\Pr((X, Y) \in C\} = \int_C \int \frac{1}{29{,}204}\, dx\, dy, \tag{3.4.1}$$

assuming that the integral exists. $\blacktriangleleft$

If one looks carefully at Eq. (3.4.1), one will notice the similarity to Eqs. (3.2.2) and (3.2.1). We formalize this connection as follows.

Definition 3.4.4 Continuous Joint Distribution/Joint p.d.f./Support. Two random variables X and Y have a *continuous joint distribution* if there exists a nonnegative function f defined over the entire xy-plane such that for every subset C of the plane,

$$\Pr[(X, Y) \in C] = \int_C \int f(x, y)\, dx\, dy,$$

if the integral exists. The function f is called the *joint probability density function* (abbreviated *joint p.d.f.*) of X and Y. The closure of the set $\{(x, y) : f(x, y) > 0\}$ is called the *support of (the distribution of)* (X, Y).

Example 3.4.5 Demands for Utilities. In Example 3.4.4, it is clear from Eq. (3.4.1) that the joint p.d.f. of X and Y is the function

$$f(x, y) = \begin{cases} \dfrac{1}{29{,}204} & \text{for } 4 \le x \le 200 \text{ and } 1 \le y \le 150, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4.2}$$

◀

It is clear from Definition 3.4.4 that the joint p.d.f. of two random variables characterizes their joint distribution. The following result is also straightforward.

Theorem 3.4.3 A joint p.d.f. must satisfy the following two conditions:

$$f(x, y) \ge 0 \quad \text{for } -\infty < x < \infty \text{ and } -\infty < y < \infty,$$

and

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

■

Any function that satisfies the two displayed formulas in Theorem 3.4.3 is the joint p.d.f. for some probability distribution.

An example of the graph of a joint p.d.f. is presented in Fig. 3.11. The total volume beneath the surface $z = f(x, y)$ and above the xy-plane must be 1. The probability that the pair (X, Y) will belong to the rectangle C is equal to the volume of the solid figure with base A shown in Fig. 3.11. The top of this solid figure is formed by the surface $z = f(x, y)$.

In Sec. 3.5, we will show that if X and Y have a continuous joint distribution, then X and Y each have a continuous distribution when considered separately. This seems reasonable intutively. However, the converse of this statement is not true, and the following result helps to show why.

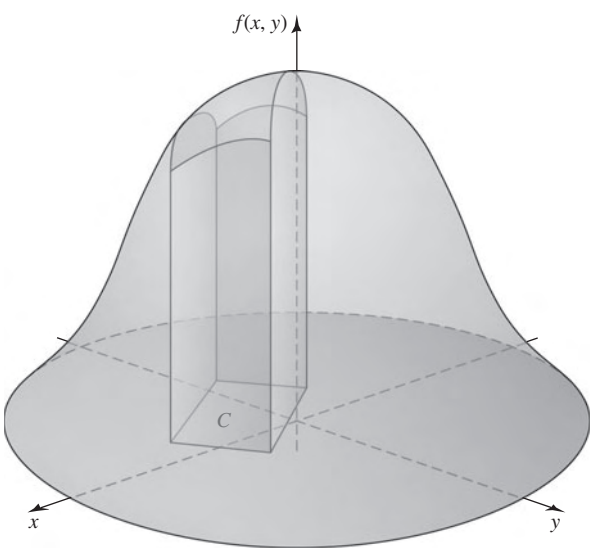

Figure 3.11 An example of a joint p.d.f.

Theorem 3.4.4 For every continuous joint distribution on the xy-plane, the following two statements hold:

i. Every individual point, and every infinite sequence of points, in the xy-plane has probability 0.

ii. Let f be a continuous function of one real variable defined on a (possibly unbounded) interval (a, b). The sets $\{(x, y) : y = f(x), a < x < b\}$ and $\{(x, y) : x = f(y), a < y < b\}$ have probability 0.

Proof According to Definition 3.4.4, the probability that a continuous joint distribution assigns to a specified region of the xy-plane can be found by integrating the joint p.d.f. $f(x, y)$ over that region, if the integral exists. If the region is a single point, the integral will be 0. By Axiom 3 of probability, the probability for any countable collection of points must also be 0. The integral of a function of two variables over the graph of a continuous function in the xy-plane is also 0. ■

Example 3.4.6 Not a Continuous Joint Distribution. It follows from (ii) of Theorem 3.4.4 that the probability that (X, Y) will lie on each specified straight line in the plane is 0. If X has a continuous distribution and if $Y = X$, then both X and Y have continuous distributions, but the probability is 1 that (X, Y) lies on the straight line $y = x$. Hence, X and Y cannot have a continuous joint distribution. ◀

Example 3.4.7 Calculating a Normalizing Constant. Suppose that the joint p.d.f. of X and Y is specified as follows:

$$f(x, y) = \begin{cases} cx^2y & \text{for } x^2 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the value of the constant c.

The support S of (X, Y) is sketched in Fig. 3.12. Since $f(x, y) = 0$ outside S, it follows that

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy &= \int_S\int f(x, y)\,dx\,dy \\ &= \int_{-1}^{1}\int_{x^2}^{1} cx^2y\,dy\,dx = \frac{4}{21}c. \end{aligned} \tag{3.4.3}$$

Since the value of this integral must be 1, the value of c must be 21/4.

The limits of integration on the last integral in (3.4.3) were determined as follows. We have our choice of whether to integrate x or y as the inner integral, and we chose y. So, we must find, for each x, the interval of y values over which to integrate. From Fig. 3.12, we see that, for each x, y runs from the curve where $y = x^2$ to the line where $y = 1$. The interval of x values for the outer integral is from -1 to 1 according to Fig. 3.12. If we had chosen to integrate x on the inside, then for each y, we see that x runs from $-\sqrt{y}$ to $\sqrt{y}$, while y runs from 0 to 1. The final answer would have been the same. ◀

Example 3.4.8 Calculating Probabilities from a Joint p.d.f. For the joint distribution in Example 3.4.7, we shall now determine the value of $\Pr(X \ge Y)$.

The subset S_0 of S where $x \ge y$ is sketched in Fig. 3.13. Hence,

$$\Pr(X \ge Y) = \int_{S_0}\int f(x, y)\,dx\,dy = \int_0^1\int_{x^2}^{x} \frac{21}{4}x^2y\,dy\,dx = \frac{3}{20}.$$ ◀

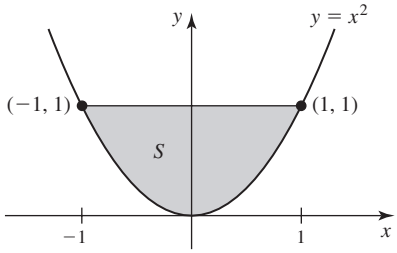

Figure 3.12 The support S of (X, Y) in Example 3.4.8.

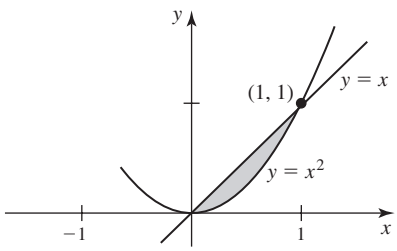

Figure 3.13 The subset S_0 of the support S where $x \geq y$ in Example 3.4.8.

Example 3.4.9

Determining a Joint p.d.f. by Geometric Methods. Suppose that a point (X, Y) is selected at random from inside the circle $x^2 + y^2 \leq 9$. We shall determine the joint p.d.f. of X and Y.

The support of (X, Y) is the set S of points on and inside the circle $x^2 + y^2 \leq 9$. The statement that the point (X, Y) is selected at random from inside the circle is interpreted to mean that the joint p.d.f. of X and Y is constant over S and is 0 outside S. Thus,

$$f(x, y) = \begin{cases} c & \text{for } (x, y) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

We must have

$$\int_S \int f(x, y)\, dx\, dy = c \times (\text{area of } S) = 1.$$

Since the area of the circle S is 9π, the value of the constant c must be $1/(9\pi)$. ◀

Mixed Bivariate Distributions

Example 3.4.10

A Clinical Trial. Consider a clinical trial (such as the one described in Example 2.1.12) in which each patient with depression receives a treatment and is followed to see whether they have a relapse into depression. Let X be the indicator of whether or not the first patient is a "success" (no relapse). That is $X = 1$ if the patient does not relapse and $X = 0$ if the patient relapses. Also, let P be the proportion of patients who have no replapse among all patients who might receive the treatment. It is clear that X must have a discrete distribution, but it might be sensible to think of P as a continuous random variable taking its value anywhere in the interval [0, 1]. Even though X and P can have neither a joint discrete distribution nor a joint continuous distribution, we can still be interested in the joint distribution of X and P. ◀

Prior to Example 3.4.10, we have discussed bivariate distributions that were either discrete or continuous. Occasionally, one must consider a mixed bivariate distribution in which one of the random variables is discrete and the other is continuous. We shall use a function $f(x, y)$ to characterize such a joint distribution in much the same way that we use a joint p.f. to characterize a discrete joint distribution or a joint p.d.f. to characterize a continuous joint distribution.

Definition 3.4.5 Joint p.f./p.d.f. Let X and Y be random variables such that X is discrete and Y is continuous. Suppose that there is a function $f(x, y)$ defined on the xy-plane such that, for every pair A and B of subsets of the real numbers,

$$\Pr(X \in A \text{ and } Y \in B) = \int_B \sum_{x \in A} f(x, y)dy, \tag{3.4.4}$$

if the integral exists. Then the function f is called the *joint p.f./p.d.f.* of X and Y.

Clearly, Definition 3.4.5 can be modified in an obvious way if Y is discrete and X is continuous. Every joint p.f./p.d.f. must satisfy two conditions. If X is the discrete random variable with possible values $x_1, x_2, \ldots$ and Y is the continuous random variable, then $f(x, y) \geq 0$ for all x, y and

$$\int_{-\infty}^{\infty} \sum_{i=1}^{\infty} f(x_i, y)dy = 1. \tag{3.4.5}$$

Because f is nonnegative, the sum and integral in Eqs. (3.4.4) and (3.4.5) can be done in whichever order is more convenient.

Note: Probabilities of More General Sets. For a general set C of pairs of real numbers, we can compute $\Pr((X, Y) \in C)$ using the joint p.f./p.d.f. of X and Y. For each x, let $C_x = \{y : (x, y) \in C\}$. Then

$$\Pr((X, Y) \in C) = \sum_{\text{All } x} \int_{C_x} f(x, y)dy,$$

if all of the integrals exist. Alternatively, for each y, define $C^y = \{x : (x, y) \in C\}$, and then

$$\Pr((X, Y) \in C) = \int_{-\infty}^{\infty} \left[\sum_{x \in C^y} f(x, y) \right] dy,$$

if the integral exists.

Example 3.4.11 A Joint p.f./p.d.f. Suppose that the joint p.f./p.d.f. of X and Y is

$$f(x, y) = \frac{xy^{x-1}}{3}, \quad \text{for } x = 1, 2, 3 \text{ and } 0 < y < 1.$$

We should check to make sure that this function satisfies (3.4.5). It is easier to integrate over the y values first, so we compute

$$\sum_{x=1}^{3} \int_0^1 \frac{xy^{x-1}}{3} dy = \sum_{x=1}^{3} \frac{1}{3} = 1.$$

Suppose that we wish to compute the probability that $Y \geq 1/2$ and $X \geq 2$. That is, we want $\Pr(X \in A \text{ and } Y \in B)$ with $A = [2, \infty)$ and $B = [1/2, \infty)$. So, we apply Eq. (3.4.4)

to get the probability

$$\sum_{x=2}^{3}\int_{1/2}^{1}\frac{xy^{x-1}}{3}dy=\sum_{x=2}^{3}\left(\frac{1-(1/2)^x}{3}\right)=0.5417.$$

For illustration, we shall compute the sum and integral in the other order also. For each $y\in[1/2,\,1)$, $\sum_{x=2}^{3}f(x,\,y)=2y/3+y^2$. For $y\geq 1/2$, the sum is 0. So, the probability is

$$\int_{1/2}^{1}\left[\frac{2}{3}y+y^2\right]dy=\frac{1}{3}\left[1-\left(\frac{1}{2}\right)^2\right]+\frac{1}{3}\left[1-\left(\frac{1}{2}\right)^3\right]=0.5417.$$ ◀

Example 3.4.12

A Clinical Trial. A possible joint p.f./p.d.f. for X and P in Example 3.4.10 is

$$f(x,\,p)=p^x(1-p)^{1-x},\quad \text{for } x=0,\,1 \text{ and } 0<p<1.$$

Here, X is discrete and P is continuous. The function f is nonnegative, and the reader should be able to demonstrate that it satisfies (3.4.5). Suppose that we wish to compute $\Pr(X\leq 0 \text{ and } P\leq 1/2)$. This can be computed as

$$\int_{0}^{1/2}(1-p)dp=-\frac{1}{2}[(1-1/2)^2-(1-0)^2]=\frac{3}{8}.$$

Suppose that we also wish to compute $\Pr(X=1)$. This time, we apply Eq. (3.4.4) with $A=\{1\}$ and $B=(0,\,1)$. In this case,

$$\Pr(X=1)=\int_{0}^{1}p\,dp=\frac{1}{2}.$$ ◀

A more complicated type of joint distribution can also arise in a practical problem.

Example 3.4.13

A Complicated Joint Distribution. Suppose that X and Y are the times at which two specific components in an electronic system fail. There might be a certain probability p $(0<p<1)$ that the two components will fail at the same time and a certain probability $1-p$ that they will fail at different times. Furthermore, if they fail at the same time, then their common failure time might be distributed according to a certain p.d.f. $f(x)$; if they fail at different times, then these times might be distributed according to a certain joint p.d.f. $g(x,\,y)$.

The joint distribution of X and Y in this example is not continuous, because there is positive probability p that $(X,\,Y)$ will lie on the line $x=y$. Nor does the joint distribution have a joint p.f./p.d.f. or any other simple function to describe it. There are ways to deal with such joint distributions, but we shall not discuss them in this text. ◀

Bivariate Cumulative Distribution Functions

The first calculation in Example 3.4.12, namely, $\Pr(X\leq 0 \text{ and } Y\leq 1/2)$, is a generalization of the calculation of a c.d.f. to a bivariate distribution. We formalize the generalization as follows.

Definition 3.4.6

Joint (Cumulative) Distribution Function/c.d.f. The *joint distribution function* or *joint cumulative distribution function* (*joint c.d.f.*) of two random variables X and Y is

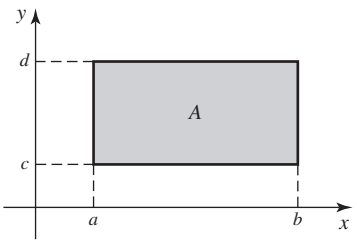

Figure 3.14 The probability of a rectangle.

defined as the function F such that for all values of x and y ($-\infty < x < \infty$ and $-\infty < y < \infty$),

$$F(x, y) = \Pr(X \le x \text{ and } Y \le y).$$

It is clear from Definition 3.4.6 that $F(x, y)$ is monotone increasing in x for each fixed y and is monotone increasing in y for each fixed x.

If the joint c.d.f. of two arbitrary random variables X and Y is F, then the probability that the pair (X, Y) will lie in a specified rectangle in the xy-plane can be found from F as follows: For given numbers $a < b$ and $c < d$,

$$\begin{aligned}
&\Pr(a < X \le b \text{ and } c < Y \le d)\\
&\quad = \Pr(a < X \le b \text{ and } Y \le d) - \Pr(a < X \le b \text{ and } Y \le c)\\
&\quad = [\Pr(X \le b \text{ and } Y \le d) - \Pr(X \le a \text{ and } Y \le d)]\\
&\qquad\quad -[\Pr(X \le b \text{ and } Y \le c) - \Pr(X \le a \text{ and } Y \le c)]\\
&\quad = F(b, d) - F(a, d) - F(b, c) + F(a, c).
\end{aligned} \tag{3.4.6}$$

Hence, the probability of the rectangle C sketched in Fig. 3.14 is given by the combination of values of F just derived. It should be noted that two sides of the rectangle are included in the set C and the other two sides are excluded. Thus, if there are points or line segments on the boundary of C that have positive probability, it is important to distinguish between the weak inequalities and the strict inequalities in Eq. (3.4.6).

Theorem 3.4.5 Let X and Y have a joint c.d.f. F. The c.d.f. F_1 of just the single random variable X can be derived from the joint c.d.f. F as $F_1(x) = \lim_{y\to\infty} F(x, y)$. Similarly, the c.d.f. F_2 of Y equals $F_2(y) = \lim_{x\to\infty} F(x, y)$, for $0 < y < \infty$.

Proof We prove the claim about F_1 as the claim about F_2 is similar. Let $-\infty < x < \infty$. Define

$$\begin{aligned}
B_0 &= \{X \le x \text{ and } Y \le 0\},\\
B_n &= \{X \le x \text{ and } n - 1 < Y \le n\}, \quad \text{for } n = 1, 2, \ldots,\\
A_m &= \bigcup_{n=0}^{m} B_n, \quad \text{for } m = 1, 2, \ldots.
\end{aligned}$$

Then $\{X \le x\} = \bigcup_{n=-0}^{\infty} B_n$, and $A_m = \{X \le x \text{ and } Y \le m\}$ for $m = 1, 2, \ldots$. It follows that $\Pr(A_m) = F(x, m)$ for each m. Also,

$$F_1(x) = \Pr(X \le x) = \Pr\left(\bigcup_{n=1}^{\infty} B_n\right)$$
$$= \sum_{n=0}^{\infty} \Pr(B_n) = \lim_{m\to\infty} \Pr(A_m)$$
$$= \lim_{m\to\infty} F(x, m) = \lim_{y\to\infty} F(x, y),$$

where the third equality follows from countable additivity and the fact that the B_n events are disjoint, and the last equality follows from the fact that $F(x, y)$ is monotone increasing in y for each fixed x. ■

Other relationships involving the univariate distribution of X, the univariate distribution of Y, and their joint bivariate distribution will be presented in the next section.

Finally, if X and Y have a continuous joint distribution with joint p.d.f. f, then the joint c.d.f. at (x, y) is

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(r, s)\, dr\, ds.$$

Here, the symbols r and s are used simply as dummy variables of integration. The joint p.d.f. can be derived from the joint c.d.f. by using the relations

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F(x, y)}{\partial y \partial x}$$

at every point (x, y) at which these second-order derivatives exist.

Example 3.4.14 Determining a Joint p.d.f. from a Joint c.d.f. Suppose that X and Y are random variables that take values only in the intervals $0 \le X \le 2$ and $0 \le Y \le 2$. Suppose also that the joint c.d.f. of X and Y, for $0 \le x \le 2$ and $0 \le y \le 2$, is as follows:

$$F(x, y) = \frac{1}{16} xy(x + y). \tag{3.4.7}$$

We shall first determine the c.d.f. F_1 of just the random variable X and then determine the joint p.d.f. f of X and Y.

The value of $F(x, y)$ at any point (x, y) in the xy-plane that does not represent a pair of possible values of X and Y can be calculated from (3.4.7) and the fact that $F(x, y) = \Pr(X \le x \text{ and } Y \le y)$. Thus, if either $x < 0$ or $y < 0$, then $F(x, y) = 0$. If both $x > 2$ and $y > 2$, then $F(x, y) = 1$. If $0 \le x \le 2$ and $y > 2$, then $F(x, y) = F(x, 2)$, and it follows from Eq. (3.4.7) that

$$F(x, y) = \frac{1}{8} x(x + 2).$$

Similarly, if $0 \le y \le 2$ and $x > 2$, then

$$F(x, y) = \frac{1}{8} y(y + 2).$$

The function $F(x, y)$ has now been specified for every point in the xy-plane.

By letting $y \to \infty$, we find that the c.d.f. of just the random variable X is

$$F_1(x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{8} x(x + 2) & \text{for } 0 \le x \le 2, \\ 1 & \text{for } x > 2. \end{cases}$$

Furthermore, for $0 < x < 2$ and $0 < y < 2$,

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{1}{8}(x + y).$$

Also, if $x < 0$, $y < 0$, $x > 2$, or $y > 2$, then

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = 0.$$

Hence, the joint p.d.f. of X and Y is

$$f(x, y) = \begin{cases} \frac{1}{8}(x + y) & \text{for } 0 < x < 2 \text{ and } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Example 3.4.15

Demands for Utilities. We can compute the joint c.d.f. for water and electric demand in Example 3.4.4 by using the joint p.d.f. that was given in Eq. (3.4.2). If either $x \le 4$ or $y \le 1$, then $F(x, y) = 0$ because either $X \le x$ or $Y \le y$ would be impossible. Similarly, if both $x \ge 200$ and $y \ge 150$, $F(x, y) = 1$ because both $X \le x$ and $Y \le y$ would be sure events. For other values of x and y, we compute

$$F(x, y) = \begin{cases} \displaystyle\int_4^x \int_1^y \frac{1}{29{,}204} dy dx = \frac{xy}{29{,}204} & \text{for } 4 \le x \le 200, 1 \le y \le 150, \\ \displaystyle\int_4^x \int_1^{150} \frac{1}{29{,}204} dy dx = \frac{x}{196} & \text{for } 4 \le x \le 200, y > 150, \\ \displaystyle\int_4^{200} \int_1^y \frac{1}{29{,}204} dy dx = \frac{y}{149} & \text{for } x > 200, 1 \le y \le 150. \end{cases}$$

The reason that we need three cases in the formula for $F(x, y)$ is that the joint p.d.f. in Eq. (3.4.2) drops to 0 when x crosses above 200 or when y crosses above 150; hence, we never want to integrate 1/29,204 beyond $x = 200$ or beyond $y = 150$. If one takes the limit as $y \to \infty$ of $F(x, y)$ (for fixed $4 \le x \le 200$), one gets the second case in the formula above, which then is the c.d.f. of X, $F_1(x)$. Similarly, if one takes the $\lim_{x \to \infty} F(x, y)$ (for fixed $1 \le y \le 150$), one gets the third case in the formula, which then is the c.d.f. of Y, $F_2(y)$. ◀

Summary

The joint c.d.f. of two random variables X and Y is $F(x, y) = \Pr(X \le x \text{ and } Y \le y)$. The joint p.d.f. of two continuous random variables is a nonnegative function f such that the probability of the pair (X, Y) being in a set C is the integral of $f(x, y)$ over the set C, if the integral exists. The joint p.d.f. is also the second mixed partial derivative of the joint c.d.f. with respect to both variables. The joint p.f. of two discrete random variables is a nonnegative function f such that the probability of the pair (X, Y) being in a set C is the sum of $f(x, y)$ over all points in C. A joint p.f. can be strictly positive at countably many pairs (x, y) at most. The joint p.f./p.d.f. of a discrete random variable X and a continuous random variable Y is a nonnegative function f such that the probability of the pair (X, Y) being in a set C is obtained by summing $f(x, y)$ over all x such that $(x, y) \in C$ for each y and then integrating the resulting function of y.

Exercises

1. Suppose that the joint p.d.f. of a pair of random variables (X, Y) is constant on the rectangle where $0 \le x \le 2$ and $0 \le y \le 1$, and suppose that the p.d.f. is 0 off of this rectangle.

a. Find the constant value of the p.d.f. on the rectangle.

b. Find $\Pr(X \ge Y)$.

2. Suppose that in an electric display sign there are three light bulbs in the first row and four light bulbs in the second row. Let X denote the number of bulbs in the first row that will be burned out at a specified time t, and let Y denote the number of bulbs in the second row that will be burned out at the same time t. Suppose that the joint p.f. of X and Y is as specified in the following table:

	Y				
X	0	1	2	3	4
0	0.08	0.07	0.06	0.01	0.01
1	0.06	0.10	0.12	0.05	0.02
2	0.05	0.06	0.09	0.04	0.03
3	0.02	0.03	0.03	0.03	0.04

Determine each of the following probabilities:

a. $\Pr(X = 2)$ **b.** $\Pr(Y \ge 2)$

c. $\Pr(X \le 2 \text{ and } Y \le 2)$ **d.** $\Pr(X = Y)$

e. $\Pr(X > Y)$

3. Suppose that X and Y have a discrete joint distribution for which the joint p.f. is defined as follows:

$$f(x, y) = \begin{cases} c|x + y| & \text{for } x = -2, -1, 0, 1, 2 \text{ and} \\ & \quad y = -2, -1, 0, 1, 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c; **(b)** $\Pr(X = 0 \text{ and } Y = -2)$; **(c)** $\Pr(X = 1)$; **(d)** $\Pr(|X - Y| \le 1)$.

4. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is defined as follows:

$$f(x, y) = \begin{cases} cy^2 & \text{for } 0 \le x \le 2 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c; **(b)** $\Pr(X + Y > 2)$; **(c)** $\Pr(Y < 1/2)$; **(d)** $\Pr(X \le 1)$; **(e)** $\Pr(X = 3Y)$.

5. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} c(x^2 + y) & \text{for } 0 \le y \le 1 - x^2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c; **(b)** $\Pr(0 \le X \le 1/2)$; **(c)** $\Pr(Y \le X + 1)$; **(d)** $\Pr(Y = X^2)$.

6. Suppose that a point (X, Y) is chosen at random from the region S in the xy-plane containing all points (x, y) such that $x \ge 0$, $y \ge 0$, and $4y + x \le 4$.

a. Determine the joint p.d.f. of X and Y.

b. Suppose that S_0 is a subset of the region S having area α and determine $\Pr[(X, Y) \in S_0]$.

7. Suppose that a point (X, Y) is to be chosen from the square S in the xy-plane containing all points (x, y) such that $0 \le x \le 1$ and $0 \le y \le 1$. Suppose that the probability that the chosen point will be the corner $(0, 0)$ is 0.1, the probability that it will be the corner $(1, 0)$ is 0.2, the probability that it will be the corner $(0, 1)$ is 0.4, and the probability that it will be the corner $(1, 1)$ is 0.1. Suppose also that if the chosen point is not one of the four corners of the square, then it will be an interior point of the square and will be chosen according to a constant p.d.f. over the interior of the square. Determine **(a)** $\Pr(X \le 1/4)$ and **(b)** $\Pr(X + Y \le 1)$.

8. Suppose that X and Y are random variables such that (X, Y) must belong to the rectangle in the xy-plane containing all points (x, y) for which $0 \le x \le 3$ and $0 \le y \le 4$. Suppose also that the joint c.d.f. of X and Y at every point (x, y) in this rectangle is specified as follows:

$$F(x, y) = \frac{1}{156} xy(x^2 + y).$$

Determine **(a)** $\Pr(1 \le X \le 2 \text{ and } 1 \le Y \le 2)$; **(b)** $\Pr(2 \le X \le 4 \text{ and } 2 \le Y \le 4)$; **(c)** the c.d.f. of Y; **(d)** the joint p.d.f. of X and Y; **(e)** $\Pr(Y \le X)$.

9. In Example 3.4.5, compute the probability that water demand X is greater than electric demand Y.

10. Let Y be the rate (calls per hour) at which calls arrive at a switchboard. Let X be the number of calls during a two-hour period. A popular choice of joint p.f./p.d.f. for (X, Y) in this example would be one like

$$f(x, y) = \begin{cases} \frac{(2y)^x}{x!} e^{-3y} & \text{if } y > 0 \text{ and } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

a. Verify that f is a joint p.f./p.d.f. *Hint:* First, sum over the x values using the well-known formula for the power series expansion of e^{2y}.

b. Find $\Pr(X = 0)$.

11. Consider the clinical trial of depression drugs in Example 2.1.4. Suppose that a patient is selected at random from the 150 patients in that study and we record Y, an

Table 3.3 Proportions in clinical depression study for Exercise 11

	Treatment group (Y)			
Response (X)	Imipramine (1)	Lithium (2)	Combination (3)	Placebo (4)
Relapse (0)	0.120	0.087	0.146	0.160
No relapse (1)	0.147	0.166	0.107	0.067

indicator of the treatment group for that patient, and X, an indicator of whether or not the patient relapsed. Table 3.3 contains the joint p.f. of X and Y.

a. Calculate the probability that a patient selected at random from this study used Lithium (either alone or in combination with Imipramine) and did not relapse.

b. Calculate the probability that the patient had a relapse (without regard to the treatment group).

3.5 Marginal Distributions

Earlier in this chapter, we introduced distributions for random variables, and in Sec. 3.4 we discussed a generalization to joint distributions of two random variables simultaneously. Often, we start with a joint distribution of two random variables and we then want to find the distribution of just one of them. The distribution of one random variable X computed from a joint distribution is also called the marginal distribution of X. Each random variable will have a marginal c.d.f. as well as a marginal p.d.f. or p.f. We also introduce the concept of independent random variables, which is a natural generalization of independent events.

Deriving a Marginal p.f. or a Marginal p.d.f.

We have seen in Theorem 3.4.5 that if the joint c.d.f. F of two random variables X and Y is known, then the c.d.f. F_1 of the random variable X can be derived from F. We saw an example of this derivation in Example 3.4.15. If X has a continuous distribution, we can also derive the p.d.f. of X from the joint distribution.

Example 3.5.1 Demands for Utilities. Look carefully at the formula for $F(x, y)$ in Example 3.4.15, specifically the last two branches that we identified as $F_1(x)$ and $F_2(y)$, the c.d.f.'s of the two individual random variables X and Y. It is apparent from those two formulas and Theorem 3.3.5 that the p.d.f. of X alone is

$$f_1(x) = \begin{cases} \frac{1}{196} & \text{for } 4 \le x \le 200, \\ 0 & \text{otherwise,} \end{cases}$$

which matches what we already found in Example 3.2.1. Similarly, the p.d.f. of Y alone is

$$f_2(y) = \begin{cases} \frac{1}{149} & \text{for } 1 \le y \le 150, \\ 0 & \text{otherwise.} \end{cases}$$

◀

The ideas employed in Example 3.5.1 lead to the following definition.

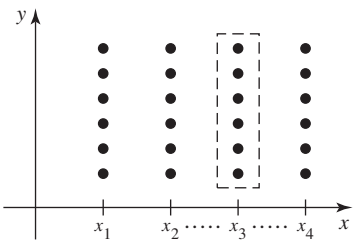

Figure 3.15 Computing $f_1(x)$ from the joint p.f.

Definition 3.5.1 Marginal c.d.f./p.f./p.d.f. Suppose that X and Y have a joint distribution. The c.d.f. of X derived by Theorem 3.4.5 is called the *marginal c.d.f.* of X. Similarly, the p.f. or p.d.f. of X associated with the marginal c.d.f. of X is called the *marginal p.f.* or *marginal p.d.f.* of X.

To obtain a specific formula for the marginal p.f. or marginal p.d.f., we start with a discrete joint distribution.

Theorem 3.5.1 If X and Y have a discrete joint distribution for which the joint p.f. is f, then the marginal p.f. f_1 of X is

$$f_1(x) = \sum_{\text{All } y} f(x, y). \tag{3.5.1}$$

Similarly, the marginal p.f. f_2 of Y is $f_2(y) = \sum_{\text{All } x} f(x, y)$.

Proof We prove the result for f_1, as the proof for f_2 is similar. We illustrate the proof in Fig. 3.15. In that figure, the set of points in the dashed box is the set of pairs with first coordinate x. The event $\{X = x\}$ can be expressed as the union of the events represented by the pairs in the dashed box, namely, $B_y = \{X = x \text{ and } Y = y\}$ for all possible y. The B_y events are disjoint and $\Pr(B_y) = f(x, y)$. Since $\Pr(X = x) = \sum_{\text{All } y} \Pr(B_y)$, Eq. (3.5.1) holds. ■

Example 3.5.2 Deriving a Marginal p.f. from a Table of Probabilities. Suppose that X and Y are the random variables in Example 3.4.3 on page 119. These are respectively the numbers of cars and televisions owned by a radomly selected household in a certain suburban area. Table 3.2 on page 119 gives their joint p.f., and we repeat that table in Table 3.4 together with row and column totals added to the margins.

The marginal p.f. f_1 of X can be read from the row totals of Table 3.4. The numbers were obtained by summing the values in each row of this table from the four columns in the central part of the table (those labeled $y = 1, 2, 3, 4$). In this way, it is found that $f_1(1) = 0.2$, $f_1(2) = 0.6$, $f_1(3) = 0.2$, and $f_1(x) = 0$ for all other values of x. This marginal p.f. gives the probabilities that a randomly selected household owns 1, 2, or 3 cars. Similarly, the marginal p.f. f_2 of Y, the probabilities that a household owns 1, 2, 3, or 4 televisions, can be read from the column totals. These numbers were obtained by adding the numbers in each of the columns from the three rows in the central part of the table (those labeled $x = 1, 2, 3$.) ◀

The name *marginal distribution* derives from the fact that the marginal distributions are the totals that appear in the margins of tables like Table 3.4.

If X and Y have a continuous joint distribution for which the joint p.d.f. is f, then the marginal p.d.f. f_1 of X is again determined in the manner shown in Eq. (3.5.1), but

Table 3.4 Joint p.f. $f(x, y)$ with marginal p.f.'s for Example 3.5.2

	y				
x	1	2	3	4	Total
1	0.1	0	0.1	0	0.2
2	0.3	0	0.1	0.2	0.6
3	0	0.2	0	0	0.2
Total	0.4	0.2	0.2	0.2	1.0

the sum over all possible values of Y is now replaced by the integral over all possible values of Y.

Theorem 3.5.2 If X and Y have a continuous joint distribution with joint p.d.f. f, then the marginal p.d.f. f_1 of X is

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{for } -\infty < x < \infty. \tag{3.5.2}$$

Similarly, the marginal p.d.f. f_2 of Y is

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx \quad \text{for } -\infty < y < \infty. \tag{3.5.3}$$

Proof We prove (3.5.2) as the proof of (3.5.3) is similar. For each x, $\Pr(X \le x)$ can be written as $\Pr((X, Y) \in C)$, where $C = \{(r, s)) : r \le x\}$. We can compute this probability directly from the joint p.d.f. of X and Y as

$$\begin{aligned} \Pr((X, Y) \in C) &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(r, s) ds dr \\ &= \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f(r, s) ds \right] dr \end{aligned} \tag{3.5.4}$$

The inner integral in the last expression of Eq. (3.5.4) is a function of r and it can easily be recognized as $f_1(r)$, where f_1 is defined in Eq. (3.5.2). It follows that $\Pr(X \le x) = \int_{-\infty}^{x} f_1(r) dr$, so f_1 is the marginal p.d.f. of X. ■

Example 3.5.3 Deriving a Marginal p.d.f. Suppose that the joint p.d.f. of X and Y is as specified in Example 3.4.8, namely,

$$f(x, y) = \begin{cases} \frac{21}{4} x^2 y & \text{for } x^2 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The set S of points (x, y) for which $f(x, y) > 0$ is sketched in Fig. 3.16. We shall determine first the marginal p.d.f. f_1 of X and then the marginal p.d.f. f_2 of Y.

It can be seen from Fig. 3.16 that X cannot take any value outside the interval $[-1, 1]$. Therefore, $f_1(x) = 0$ for $x < -1$ or $x > 1$. Furthermore, for $-1 \le x \le 1$, it is seen from Fig. 3.16 that $f(x, y) = 0$ unless $x^2 \le y \le 1$. Therefore, for $-1 \le x \le 1$,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{x^2}^{1} \left(\frac{21}{4} \right) x^2 y\, dy = \left(\frac{21}{8} \right) x^2 (1 - x^4).$$

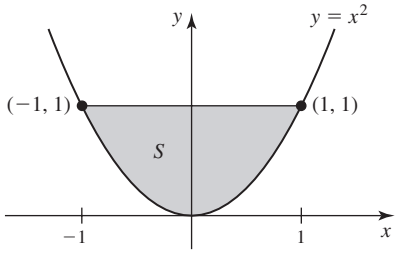

Figure 3.16 The set S where $f(x, y) > 0$ in Example 3.5.3.

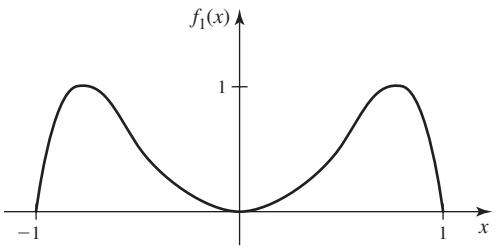

Figure 3.17 The marginal p.d.f. of X in Example 3.5.3.

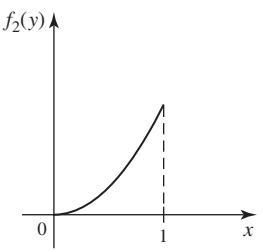

Figure 3.18 The marginal p.d.f. of Y in Example 3.5.3.

This marginal p.d.f. of X is sketched in Fig. 3.17.

Next, it can be seen from Fig. 3.16 that Y cannot take any value outside the interval $[0, 1]$. Therefore, $f_2(y) = 0$ for $y < 0$ or $y > 1$. Furthermore, for $0 \le y \le 1$, it is seen from Fig. 3.12 that $f(x, y) = 0$ unless $-\sqrt{y} \le x \le \sqrt{y}$. Therefore, for $0 \le y \le 1$,

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{-\sqrt{y}}^{\sqrt{y}} \left(\frac{21}{4}\right) x^2 y\, dx = \left(\frac{7}{2}\right) y^{5/2}.$$

This marginal p.d.f. of Y is sketched in Fig. 3.18. ◀

If X has a discrete distribution and Y has a continuous distribution, we can derive the marginal p.f. of X and the marginal p.d.f. of Y from the joint p.f./p.d.f. in the same ways that we derived a marginal p.f. or a marginal p.d.f. from a joint p.f. or a joint p.d.f. The following result can be proven by combining the techniques used in the proofs of Theorems 3.5.1 and 3.5.2.

Theorem 3.5.3 Let f be the joint p.f./p.d.f. of X and Y, with X discrete and Y continuous. Then the marginal p.f. of X is

$$f_1(x) = \Pr(X = x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad \text{for all } x,$$

and the marginal p.d.f. of Y is

$$f_2(y) = \sum_x f(x, y), \quad \text{for } -\infty < y < \infty. \qquad \blacksquare$$

Example 3.5.4 Determining a Marginal p.f. and Marginal p.d.f. from a Joint p.f./p.d.f. Suppose that the joint p.f./p.d.f. of X and Y is as in Example 3.4.11 on page 124. The marginal p.f. of X is obtained by integrating

$$f_1(x) = \int_0^1 \frac{xy^{x-1}}{3} dy = \frac{1}{3},$$

for $x = 1, 2, 3$. The marginal p.d.f. of Y is obtained by summing

$$f_2(y) = \frac{1}{3} + \frac{2y}{3} + y^2, \quad \text{for } 0 < y < 1. \qquad \blacktriangleleft$$

Although the marginal distributions of X and Y can be derived from their joint distribution, it is not possible to reconstruct the joint distribution of X and Y from their marginal distributions without additional information. For instance, the marginal p.d.f.'s sketched in Figs. 3.17 and 3.18 reveal no information about the relationship between X and Y. In fact, by definition, the marginal distribution of X specifies probabilities for X without regard for the values of any other random variables. This property of a marginal p.d.f. can be further illustrated by another example.

Example 3.5.5 Marginal and Joint Distributions. Suppose that a penny and a nickel are each tossed n times so that every pair of sequences of tosses (n tosses in each sequence) is equally likely to occur. Consider the following two definitions of X and Y: (i) X is the number of heads obtained with the penny, and Y is the number of heads obtained with the nickel. (ii) Both X and Y are the number of heads obtained with the penny, so the random variables X and Y are actually identical.

In case (i), the marginal distribution of X and the marginal distribution of Y will be identical binomial distributions. The same pair of marginal distributions of X and Y will also be obtained in case (ii). However, the joint distribution of X and Y will not be the same in the two cases. In case (i), X and Y can take different values. Their joint p.f. is

$$f(x, y) = \begin{cases} \binom{n}{x}\binom{n}{y}\left(\frac{1}{2}\right)^{x+y} & \text{for } x = 0, 1 \ldots, n,\ y = 0, 1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

In case (ii), X and Y must take the same value, and their joint p.f. is

$$f(x, y) = \begin{cases} \binom{n}{x}\left(\frac{1}{2}\right)^{x} & \text{for } x = y = 0, 1 \ldots, n, \\ 0 & \text{otherwise.} \end{cases} \qquad \blacktriangleleft$$

Independent Random Variables

Example 3.5.6 Demands for Utilities. In Examples 3.4.15 and 3.5.1, we found the marginal c.d.f.'s of water and electric demand were, respectively,

$$F_1(x) = \begin{cases} 0 & \text{for } x < 4, \\ \frac{x}{196} & \text{for } 4 \le x \le 200, \\ 1 & \text{for } x > 200, \end{cases} \qquad F_2(y) = \begin{cases} 0 & \text{for } y < 1, \\ \frac{y}{149} & \text{for } 1 \le y \le 150, \\ 1 & \text{for } y > 150. \end{cases}$$

The product of these two functions is precisely the same as the joint c.d.f. of X and Y given in Example 3.5.1. One consequence of this fact is that, for every x and y, $\Pr(X \le x, \text{ and } Y \le y) = \Pr(X \le x)\Pr(Y \le y)$. This equation makes X and Y an example of the next definition. ◀

Definition 3.5.2 Independent Random Variables. It is said that two random variables X and Y are *independent* if, for every two sets A and B of real numbers such that $\{X \in A\}$ and $\{Y \in B\}$ are events,

$$\Pr(X \in A \text{ and } Y \in B) = \Pr(X \in A)\Pr(Y \in B). \tag{3.5.5}$$

In other words, let E be any event the occurrence or nonoccurrence of which depends only on the value of X (such as $E = \{X \in A\}$), and let D be any event the occurrence or nonoccurrence of which depends only on the value of Y (such as $D = \{Y \in B\}$). Then X and Y are independent random variables if and only if E and D are independent events for all such events E and D.

If X and Y are independent, then for all real numbers x and y, it must be true that

$$\Pr(X \le x \text{ and } Y \le y) = \Pr(X \le x)\Pr(Y \le y). \tag{3.5.6}$$

Moreover, since all probabilities for X and Y of the type appearing in Eq. (3.5.5) can be derived from probabilities of the type appearing in Eq. (3.5.6), it can be shown that if Eq. (3.5.6) is satisfied for all values of x and y, then X and Y must be independent. The proof of this statement is beyond the scope of this book and is omitted, but we summarize it as the following theorem.

Theorem 3.5.4 Let the joint c.d.f. of X and Y be F, let the marginal c.d.f. of X be F_1, and let the marginal c.d.f. of Y be F_2. Then X and Y are independent if and only if, for all real numbers x and y, $F(x, y) = F_1(x)F_2(y)$. ■

For example, the demands for water and electricity in Example 3.5.6 are independent. If one returns to Example 3.5.1, one also sees that the product of the marginal p.d.f.'s of water and electric demand equals their joint p.d.f. given in Eq. (3.4.2). This relation is characteristic of independent random variables whether discrete or continuous.

Theorem 3.5.5 Suppose that X and Y are random variables that have a joint p.f., p.d.f., or p.f./p.d.f. f. Then X and Y will be independent if and only if f can be represented in the following form for $-\infty < x < \infty$ and $-\infty < y < \infty$:

$$f(x, y) = h_1(x)h_2(y), \tag{3.5.7}$$

where h_1 is a nonnegative function of x alone and h_2 is a nonnegative function of y alone.

Proof We shall give the proof only for the case in which X is discrete and Y is continuous. The other cases are similar. For the "if" part, assume that Eq. (3.5.7) holds. Write

$$f_1(x) = \int_{-\infty}^{\infty} h_1(x)h_2(y)dy = c_1h_1(x),$$

where $c_1 = \int_{-\infty}^{\infty} h_2(y)dy$ must be finite and strictly positive, otherwise f_1 wouldn't be a p.f. So, $h_1(x) = f_1(x)/c_1$. Similarly,

$$f_2(y) = \sum_x h_1(x)h_2(y) = h_2(y)\sum_x \frac{1}{c_1}f_1(x) = \frac{1}{c_1}h_2(y).$$

So, $h_2(y) = c_1 f_2(y)$. Since $f(x, y) = h_1(x)h_2(y)$, it follows that

$$f(x, y) = \frac{f_1(x)}{c_1} c_1 f_2(y) = f_1(x) f_2(y). \tag{3.5.8}$$

Now let A and B be sets of real numbers. Assuming the integrals exist, we can write

$$\begin{aligned}\Pr(X \in A \text{ and } Y \in B) &= \sum_{x \in A} \int_B f(x, y) dy \\ &= \int_B \sum_{x \in A} f_1(x) f_2(y) dy, \\ &= \sum_{x \in A} f_1(x) \int_B f_2(y) dy,\end{aligned}$$

where the first equality is from Definition 3.4.5, the second is from Eq. (3.5.8), and the third is straightforward rearrangement. We now see that X and Y are independent according to Definition 3.5.2.

For the "only if" part, assume that X and Y are independent. Let A and B be sets of real numbers. Let f_1 be the marginal p.d.f. of X, and let f_2 be the marginal p.f. of Y. Then

$$\begin{aligned}\Pr(X \in A \text{ and } Y \in B) &= \sum_{x \in A} f_1(x) \int_B f_2(y) dy \\ &= \int_B \sum_{x \in A} f_1(x) f_2(y) dy,\end{aligned}$$

(if the integral exists) where the first equality follows from Definition 3.5.2 and the second is a straightforward rearrangement. We now see that $f_1(x) f_2(y)$ satisfies the conditions needed to be $f(x, y)$ as stated in Definition 3.4.5. ■

A simple corollary follows from Theorem 3.5.5.

Corollary 3.5.1 Two random variables X and Y are independent if and only if the following factorization is satisfied for all real numbers x and y:

$$f(x, y) = f_1(x) f_2(y). \tag{3.5.9}$$

■

As stated in Sec. 3.2 (see page 102), in a continuous distribution the values of a p.d.f. can be changed arbitrarily at any countable set of points. Therefore, for such a distribution it would be more precise to state that the random variables X and Y are independent if and only if it is possible to choose versions of f, f_1, and f_2 such that Eq. (3.5.9) is satisfied for $-\infty < x < \infty$ and $-\infty < y < \infty$.

The Meaning of Independence We have given a mathematical definition of independent random variables in Definition 3.5.2, but we have not yet given any interpretation of the concept of independent random variables. Because of the close connection between independent events and independent random variables, the interpretation of independent random variables should be closely related to the interpretation of independent events. We model two events as independent if learning that one of them occurs does not change the probability that the other one occurs. It is easiest to extend this idea to discrete random variables. Suppose that X and Y

Table 3.5 Joint p.f. $f(x, y)$ for Example 3.5.7

	y						
x	1	2	3	4	5	6	Total
0	1/24	1/24	1/24	1/24	1/24	1/24	1/4
1	1/12	1/12	1/12	1/12	1/12	1/12	1/2
2	1/24	1/24	1/24	1/24	1/24	1/24	1/4
Total	1/6	1/6	1/6	1/6	1/6	1/6	1.000

have a discrete joint distribution. If, for each y, learning that $Y = y$ does not change any of the probabilities of the events $\{X = x\}$, we would like to say that X and Y are independent. From Corollary 3.5.1 and the definition of marginal p.f., we see that indeed X and Y are independent if and only if, for each y and x such that $\Pr(Y = y) > 0$, $\Pr(X = x|Y = y) = \Pr(X = x)$, that is, learning the value of Y doesn't change any of the probabilities associated with X. When we formally define conditional distributions in Sec. 3.6, we shall see that this interpretation of independent discrete random variables extends to all bivariate distributions. In summary, if we are trying to decide whether or not to model two random variables X and Y as independent, we should think about whether we would change the distribution of X after we learned the value of Y or vice versa.

Example 3.5.7

Games of Chance. A carnival game consists of rolling a fair die, tossing a fair coin two times, and recording both outcomes. Let Y stand for the number on the die, and let X stand for the number of heads in the two tosses. It seems reasonable to believe that all of the events determined by the roll of the die are independent of all of the events determined by the flips of the coin. Hence, we can assume that X and Y are independent random variables. The marginal distribution of Y is the uniform distribution on the integers $1, \ldots, 6$, while the distribution of X is the binomial distribution with parameters 2 and 1/2. The marginal p.f.'s and the joint p.f. of X and Y are given in Table 3.5, where the joint p.f. was constructed using Eq. (3.5.9). The Total column gives the marginal p.f. f_1 of X, and the Total row gives the marginal p.f. f_2 of Y. ◀

Example 3.5.8

Determining Whether Random Variables Are Independent in a Clinical Trial. Return to the clinical trial of depression drugs in Exercise 11 of Sec. 3.4 (on page 129). In that trial, a patient is selected at random from the 150 patients in the study and we record Y, an indicator of the treatment group for that patient, and X, an indicator of whether or not the patient relapsed. Table 3.6 repeats the joint p.f. of X and Y along with the marginal distributions in the margins. We shall determine whether or not X and Y are independent.

In Eq. (3.5.9), $f(x, y)$ is the probability in the xth row and the yth column of the table, $f_1(x)$ is the number in the Total column in the xth row, and $f_2(y)$ is the number in the Total row in the yth column. It is seen in the table that $f(1, 2) = 0.087$, while $f_1(1) = 0.513$, and $f_2(1) = 0.253$. Hence, $f(1, 2) \neq f_1(1) f_2(1) = 0.129$. It follows that X and Y are not independent. ◀

It should be noted from Examples 3.5.7 and 3.5.8 that X and Y will be independent if and only if the rows of the table specifying their joint p.f. are proportional to

Table 3.6 Proportions marginals in Example 3.5.8

	Treatment group (Y)				
Response (X)	Imipramine (1)	Lithium (2)	Combination (3)	Placebo (4)	Total
Relapse (0)	0.120	0.087	0.146	0.160	0.513
No relapse (1)	0.147	0.166	0.107	0.067	0.487
Total	0.267	0.253	0.253	0.227	1.0

one another, or equivalently, if and only if the columns of the table are proportional to one another.

Example 3.5.9 Calculating a Probability Involving Independent Random Variables. Suppose that two measurements X and Y are made of the rainfall at a certain location on May 1 in two consecutive years. It might be reasonable, given knowledge of the history of rainfall on May 1, to treat the random variables X and Y as independent. Suppose that the p.d.f. g of each measurement is as follows:

$$g(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the value of $\Pr(X + Y \le 1)$.

Since X and Y are independent and each has the p.d.f. g, it follows from Eq. (3.5.9) that for all values of x and y the joint p.d.f. $f(x, y)$ of X and Y will be specified by the relation $f(x, y) = g(x)g(y)$. Hence,

$$f(x, y) = \begin{cases} 4xy & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The set S in the xy-plane, where $f(x, y) > 0$, and the subset S_0, where $x + y \le 1$, are sketched in Fig. 3.19. Thus,

$$\Pr(X + Y \le 1) = \int_{S_0} \int f(x, y)\, dx\, dy = \int_0^1 \int_0^{1-x} 4xy\, dy\, dx = \frac{1}{6}.$$

As a final note, if the two measurements X and Y had been made on the same day at nearby locations, then it might not make as much sense to treat them as independent, since we would expect them to be more similar to each other than to historical rainfalls. For example, if we first learn that X is small compared to historical rainfall on the date in question, we might then expect Y to be smaller than the historical distribution would suggest. ◀

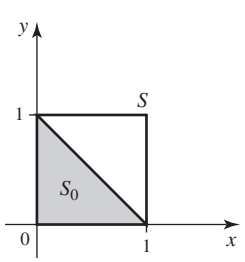

Figure 3.19 The subset S_0 where $x + y \le 1$ in Example 3.5.9.

Theorem 3.5.5 says that X and Y are independent if and only if, for all values of x and y, f can be factored into the product of an arbitrary nonnegative function of x and an arbitrary nonnegative function of y. However, it should be emphasized that, just as in Eq. (3.5.9), the factorization in Eq. (3.5.7) must be satisfied for all values of x and y ($-\infty < x < \infty$ and $-\infty < y < \infty$).

Example 3.5.10 Dependent Random Variables. Suppose that the joint p.d.f. of X and Y has the following form:

$$f(x, y) = \begin{cases} kx^2y^2 & \text{for } x^2 + y^2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall show that X and Y are not independent.

It is evident that at each point inside the circle $x^2 + y^2 \le 1$, $f(x, y)$ can be factored as in Eq. (3.5.7). However, this same factorization cannot also be satisfied at every point outside this circle. For example, $f(0.9, 0.9) = 0$, but neither $f_1(0.9) = 0$ nor $f_2(0.9) = 0$. (In Exercise 13, you can verify this feature of f_1 and f_2.)

The important feature of this example is that the values of X and Y are constrained to lie inside a circle. The joint p.d.f. of X and Y is positive inside the circle and zero outside the circle. Under these conditions, X and Y cannot be independent, because for every given value y of Y, the possible values of X will depend on y. For example, if $Y = 0$, then X can have any value such that $X^2 \le 1$; if $Y = 1/2$, then X must have a value such that $X^2 \le 3/4$. ◀

Example 3.5.10 shows that one must be careful when trying to apply Theorem 3.5.5. The situation that arose in that example will occur whenever $\{(x, y) : f(x, y) > 0\}$ has boundaries that are curved or not parallel to the coordinate axes. There is one important special case in which it is easy to check the conditions of Theorem 3.5.5. The proof is left as an exercise.

Theorem 3.5.6 Let X and Y have a continuous joint distribution. Suppose that $\{(x, y) : f(x, y) > 0\}$ is a rectangular region R (possibly unbounded) with sides (if any) parallel to the coordinate axes. Then X and Y are independent if and only if Eq. (3.5.7) holds for all $(x, y) \in R$. ■

Example 3.5.11 Verifying the Factorization of a Joint p.d.f. Suppose that the joint p.d.f. f of X and Y is as follows:

$$f(x, y) = \begin{cases} ke^{-(x+2y)} & \text{for } x \ge 0 \text{ and } y \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

where k is some constant. We shall first determine whether X and Y are independent and then determine their marginal p.d.f.'s.

In this example, $f(x, y) = 0$ outside of an unbounded rectangular region R whose sides are the lines $x = 0$ and $y = 0$. Furthermore, at each point inside R, $f(x, y)$ can be factored as in Eq. (3.5.7) by letting $h_1(x) = ke^{-x}$ and $h_2(y) = e^{-2y}$. Therefore, X and Y are independent.

It follows that in this case, except for constant factors, $h_1(x)$ for $x \ge 0$ and $h_2(y)$ for $y \ge 0$ must be the marginal p.d.f.'s of X and Y. By choosing constants that make $h_1(x)$ and $h_2(y)$ integrate to unity, we can conclude that the marginal p.d.f.'s f_1 and f_2 of X and Y must be as follows:

$$f_1(x) = \begin{cases} e^{-x} & \text{for } x \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_2(y) = \begin{cases} 2e^{-2y} & \text{for } y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

If we multiply $f_1(x)$ times $f_2(y)$ and compare the product to $f(x, y)$, we see that $k = 2$. ◀

Note: Separate Functions of Independent Random Variables Are Independent. If X and Y are independent, then $h(X)$ and $g(Y)$ are independent no matter what the functions h and g are. This is true because for every t, the event $\{h(X) \leq t\}$ can always be written as $\{X \in A\}$, where $A = \{x : h(x) \leq t\}$. Similarly, $\{g(Y) \leq u\}$ can be written as $\{Y \in B\}$, so Eq. (3.5.6) for $h(X)$ and $g(Y)$ follows from Eq. (3.5.5) for X and Y.

Summary

Let $f(x, y)$ be a joint p.f., joint p.d.f., or joint p.f./p.d.f. of two random variables X and Y. The marginal p.f. or p.d.f. of X is denoted by $f_1(x)$, and the marginal p.f. or p.d.f. of Y is denoted by $f_2(y)$. To obtain $f_1(x)$, compute $\sum_y f(x, y)$ if Y is discrete or $\int_{-\infty}^{\infty} f(x, y)\, dy$ if Y is continuous. Similarly, to obtain $f_2(y)$, compute $\sum_x f(x, y)$ if X is discrete or $\int_{-\infty}^{\infty} f(x, y)\, dx$ if X is continuous. The random variables X and Y are independent if and only if $f(x, y) = f_1(x) f_2(y)$ for *all* x and y. This is true regardless of whether X and/or Y is continuous or discrete. A sufficient condition for two continuous random variables to be independent is that $R = \{(x, y) : f(x, y) > 0\}$ be rectangular with sides parallel to the coordinate axes and that $f(x, y)$ factors into separate functions of x of y in R.

Exercises

1. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is

$$f(x, y) = \begin{cases} k & \text{for } a \leq x \leq b \text{ and } c \leq y \leq d, \\ 0 & \text{otherwise,} \end{cases}$$

where $a < b$, $c < d$, and $k > 0$. Find the marginal distributions of X and Y.

2. Suppose that X and Y have a discrete joint distribution for which the joint p.f. is defined as follows:

$$f(x, y) = \begin{cases} \frac{1}{30}(x + y) & \text{for } x = 0, 1, 2 \text{ and } y = 0, 1, 2, 3, \\ 0 & \text{otherwise.} \end{cases}$$

a. Determine the marginal p.f.'s of X and Y.

b. Are X and Y independent?

3. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is defined as follows:

$$f(x, y) = \begin{cases} \frac{3}{2}y^2 & \text{for } 0 \leq x \leq 2 \text{ and } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. Determine the marginal p.d.f.'s of X and Y.

b. Are X and Y independent?

c. Are the event $\{X < 1\}$ and the event $\{Y \geq 1/2\}$ independent?

4. Suppose that the joint p.d.f. of X and Y is as follows:

$$f(x, y) = \begin{cases} \frac{15}{4}x^2 & \text{for } 0 \leq y \leq 1 - x^2, \\ 0 & \text{otherwise.} \end{cases}$$

a. Determine the marginal p.d.f.'s of X and Y.

b. Are X and Y independent?

5. A certain drugstore has three public telephone booths. For $i = 0, 1, 2, 3$, let p_i denote the probability that exactly i telephone booths will be occupied on any Monday evening at 8:00 P.M.; and suppose that $p_0 = 0.1$, $p_1 = 0.2$, $p_2 = 0.4$, and $p_3 = 0.3$. Let X and Y denote the number of booths that will be occupied at 8:00 P.M. on two independent Monday evenings. Determine: **(a)** the joint p.f. of X and Y; **(b)** $\Pr(X = Y)$; **(c)** $\Pr(X > Y)$.

6. Suppose that in a certain drug the concentration of a particular chemical is a random variable with a continuous distribution for which the p.d.f. g is as follows:

$$g(x) = \begin{cases} \frac{3}{8}x^2 & \text{for } 0 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that the concentrations X and Y of the chemical in two separate batches of the drug are independent random variables for each of which the p.d.f. is g. Determine **(a)** the joint p.d.f. of X and Y; **(b)** $\Pr(X = Y)$; **(c)** $\Pr(X > Y)$; **(d)** $\Pr(X + Y \le 1)$.

7. Suppose that the joint p.d.f. of X and Y is as follows:

$$f(x, y) = \begin{cases} 2xe^{-y} & \text{for } 0 \le x \le 1 \text{ and } 0 < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

8. Suppose that the joint p.d.f. of X and Y is as follows:

$$f(x, y) = \begin{cases} 24xy & \text{for } x \ge 0,\ y \ge 0, \text{ and } x + y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

9. Suppose that a point (X, Y) is chosen at random from the rectangle S defined as follows:

$$S = \{(x, y) : 0 \le x \le 2 \text{ and } 1 \le y \le 4\}.$$

a. Determine the joint p.d.f. of X and Y, the marginal p.d.f. of X, and the marginal p.d.f. of Y.

b. Are X and Y independent?

10. Suppose that a point (X, Y) is chosen at random from the circle S defined as follows:

$$S = \{(x, y) : x^2 + y^2 \le 1\}.$$

a. Determine the joint p.d.f. of X and Y, the marginal p.d.f. of X, and the marginal p.d.f. of Y.

b. Are X and Y independent?

11. Suppose that two persons make an appointment to meet between 5 P.M. and 6 P.M. at a certain location, and they agree that neither person will wait more than 10 minutes for the other person. If they arrive independently at random times between 5 P.M. and 6 P.M., what is the probability that they will meet?

12. Prove Theorem 3.5.6.

13. In Example 3.5.10, verify that X and Y have the same marginal p.d.f.'s and that

$$f_1(x) = \begin{cases} 2kx^2(1 - x^2)^{3/2}/3 & \text{if } -1 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

14. For the joint p.d.f. in Example 3.4.7, determine whether or not X and Y are independent.

15. A painting process consists of two stages. In the first stage, the paint is applied, and in the second stage, a protective coat is added. Let X be the time spent on the first stage, and let Y be the time spent on the second stage. The first stage involves an inspection. If the paint fails the inspection, one must wait three minutes and apply the paint again. After a second application, there is no further inspection. The joint p.d.f. of X and Y is

$$f(x, y) = \begin{cases} \frac{1}{3} & \text{if } 1 < x < 3 \text{ and } 0 < y < 1, \\ \frac{1}{6} & \text{if } 6 < x < 8 \text{ and } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. Sketch the region where $f(x, y) > 0$. Note that it is not exactly a rectangle.

b. Find the marginal p.d.f.'s of X and Y.

c. Show that X and Y are independent.

This problem does not contradict Theorem 3.5.6. In that theorem the conditions, including that the set where $f(x, y) > 0$ be rectangular, are sufficient but not necessary.

3.6 Conditional Distributions

We generalize the concept of conditional probability to conditional distributions. Recall that distributions are just collections of probabilities of events determined by random variables. Conditional distributions will be the probabilities of events determined by some random variables conditional on events determined by other random variables. The idea is that there will typically be many random variables of interest in an applied problem. After we observe some of those random variables, we want to be able to adjust the probabilities associated with the ones that have not yet been observed. The conditional distribution of one random variable X given another Y will be the distribution that we would use for X after we learn the value of Y.

Table 3.7 Joint p.f. for Example 3.6.1

	Brand Y					
Stolen X	1	2	3	4	5	Total
0	0.129	0.298	0.161	0.280	0.108	0.976
1	0.010	0.010	0.001	0.002	0.001	0.024
Total	0.139	0.308	0.162	0.282	0.109	1.000

Discrete Conditional Distributions

Example 3.6.1

Auto Insurance. Insurance companies keep track of how likely various cars are to be stolen. Suppose that a company in a particular area computes the joint distribution of car brands and the indicator of whether the car will be stolen during a particular year that appears in Table 3.7.

We let $X = 1$ mean that a car is stolen, and we let $X = 0$ mean that the car is not stolen. We let Y take one of the values from 1 to 5 to indicate the brand of car as indicated in Table 3.7. If a customer applies for insurance for a particular brand of car, the company needs to compute the distribution of the random variable X as part of its premium determination. The insurance company might adjust their premium according to a risk factor such as likelihood of being stolen. Although, overall, the probability that a car will be stolen is 0.024, if we assume that we know the brand of car, the probability might change quite a bit. This section introduces the formal concepts for addressing this type of problem. ◀

Suppose that X and Y are two random variables having a discrete joint distribution for which the joint p.f. is f. As before, we shall let f_1 and f_2 denote the marginal p.f.'s of X and Y, respectively. After we observe that $Y = y$, the probability that the random variable X will take a particular value x is specified by the following conditional probability:

$$\Pr(X = x|Y = y) = \frac{\Pr(X = x \text{ and } Y = y)}{\Pr(Y = y)} = \frac{f(x, y)}{f_2(y)}. \tag{3.6.1}$$

In other words, if it is known that $Y = y$, then the probability that $X = x$ will be updated to the value in Eq. (3.6.1). Next, we consider the entire distribution of X after learning that $Y = y$.

Definition 3.6.1

Conditional Distribution/p.f. Let X and Y have a discrete joint distribution with joint p.f. f. Let f_2 denote the marginal p.f. of Y. For each y such that $f_2(y) > 0$, define

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)}. \tag{3.6.2}$$

Then g_1 is called the *conditional p.f. of X given Y*. The discrete distribution whose p.f. is $g_1(\cdot|y)$ is called the *conditional distribution of X given that $Y = y$*.

Table 3.8 Conditional p.f. of Y given X for Example 3.6.3

	Brand Y				
Stolen X	1	2	3	4	5
0	0.928	0.968	0.994	0.993	0.991
1	0.072	0.032	0.006	0.007	0.009

We should verify that $g_1(x|y)$ is actually a p.f. as a function of x for each y. Let y be such that $f_2(y) > 0$. Then $g_1(x|y) \geq 0$ for all x and

$$\sum_x g_1(x|y) = \frac{1}{f_2(y)} \sum_x f(x, y) = \frac{1}{f_2(y)} f_2(y) = 1.$$

Notice that we do not bother to define $g_1(x|y)$ for those y such that $f_2(y) = 0$.

Similarly, if x is a given value of X such that $f_1(x) = \Pr(X = x) > 0$, and if $g_2(y|x)$ is the *conditional p.f. of Y given that $X = x$*, then

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)}. \tag{3.6.3}$$

For each x such that $f_1(x) > 0$, the function $g_2(y|x)$ will be a p.f. as a function of y.

Example 3.6.2 Calculating a Conditional p.f. from a Joint p.f. Suppose that the joint p.f. of X and Y is as specified in Table 3.4 in Example 3.5.2. We shall determine the conditional p.f. of Y given that $X = 2$.

The marginal p.f. of X appears in the Total column of Table 3.4, so $f_1(2) = \Pr(X = 2) = 0.6$. Therefore, the conditional probability $g_2(y|2)$ that Y will take a particular value y is

$$g_2(y|2) = \frac{f(2, y)}{0.6}.$$

It should be noted that for all possible values of y, the conditional probabilities $g_2(y|2)$ must be proportional to the joint probabilities $f(2, y)$. In this example, each value of $f(2, y)$ is simply divided by the constant $f_1(2) = 0.6$ in order that the sum of the results will be equal to 1. Thus,

$$g_2(1|2) = 1/2, \quad g_2(2|2) = 0, \quad g_2(3|2) = 1/6, \quad g_2(4|2) = 1/3. \quad \blacktriangleleft$$

Example 3.6.3 Auto Insurance. Consider again the probabilities of car brands and cars being stolen in Example 3.6.1. The conditional distribution of X (being stolen) given Y (brand) is given in Table 3.8. It appears that Brand 1 is much more likely to be stolen than other cars in this area, with Brand 1 also having a significant chance of being stolen. $\blacktriangleleft$

Continuous Conditional Distributions

Example 3.6.4 Processing Times. A manufacturing process consists of two stages. The first stage takes Y minutes, and the whole process takes X minutes (which includes the first

Y minutes). Suppose that X and Y have a joint continuous distribution with joint p.d.f.

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

After we learn how much time Y that the first stage takes, we want to update our distribution for the total time X. In other words, we would like to be able to compute a conditional distribution for X given $Y = y$. We cannot argue the same way as we did with discrete joint distributions, because $\{Y = y\}$ is an event with probability 0 for all y. ◀

To facilitate the solutions of problems such as the one posed in Example 3.6.4, the concept of conditional probability will be extended by considering the definition of the conditional p.f. of X given in Eq. (3.6.2) and the analogy between a p.f. and a p.d.f.

Definition 3.6.2 Conditional p.d.f. Let X and Y have a continuous joint distribution with joint p.d.f. f and respective marginals f_1 and f_2. Let y be a value such that $f_2(y) > 0$. Then the *conditional p.d.f. g_1 of X given that $Y = y$* is defined as follows:

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)} \quad \text{for } -\infty < x < \infty. \tag{3.6.4}$$

For values of y such that $f_2(y) = 0$, we are free to define $g_1(x|y)$ however we wish, so long as $g_1(x|y)$ is a p.d.f. as a function of x.

It should be noted that Eq. (3.6.2) and Eq. (3.6.4) are identical. However, Eq. (3.6.2) was *derived* as the conditional probability that $X = x$ given that $Y = y$, whereas Eq. (3.6.4) was *defined* to be the value of the conditional p.d.f. of X given that $Y = y$. In fact, we should verify that $g_1(x|y)$ as defined above really is a p.d.f.

Theorem 3.6.1 For each y, $g_1(x|y)$ defined in Definition 3.6.2 is a p.d.f. as a function of x.

Proof If $f_2(y) = 0$, then g_1 is defined to be any p.d.f. we wish, and hence it is a p.d.f. If $f_2(y) > 0$, g_1 is defined by Eq. (3.6.4). For each such y, it is clear that $g_1(x|y) \ge 0$ for all x. Also, if $f_2(y) > 0$, then

$$\int_{-\infty}^{\infty} g_1(x|y)\, dx = \frac{\int_{-\infty}^{\infty} f(x, y)\, dx}{f_2(y)} = \frac{f_2(y)}{f_2(y)} = 1,$$

by using the formula for $f_2(y)$ in Eq. (3.5.3). ■

Example 3.6.5 Processing Times. In Example 3.6.4, Y is the time that the first stage of a process takes, while X is the total time of the two stages. We want to calculate the conditional p.d.f. of X given Y. We can calculate the marginal p.d.f. of Y as follows: For each y, the possible values of X are all $x \ge y$, so for each $y > 0$,

$$f_2(y) = \int_y^{\infty} e^{-x} dx = e^{-y},$$

and $f_2(y) = 0$ for $y < 0$. For each $y \ge 0$, the conditional p.d.f. of X given $Y = y$ is then

$$g_1(x|y) = \frac{f(x, y)}{f_2(y)} = \frac{e^{-x}}{e^{-y}} = e^{y-x}, \text{ for } x \ge y,$$

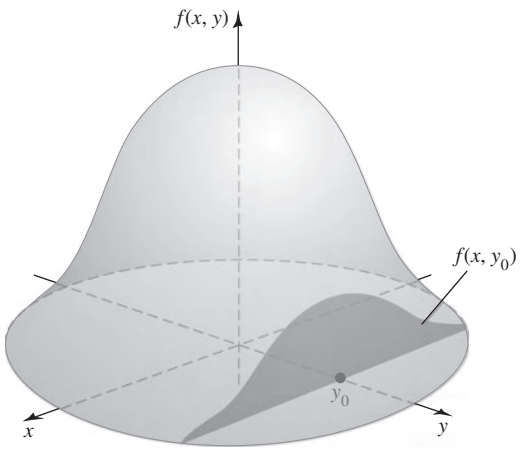

Figure 3.20 The conditional p.d.f. $g_1(x|y_0)$ is proportional to $f(x, y_0)$.

and $g_1(x|y) = 0$ for $x < y$. So, for example, if we observe $Y = 4$ and we want the conditional probability that $X \geq 9$, we compute

$$\Pr(X \geq 9|Y = 4) = \int_9^\infty e^{4-x}dx = e^{-5} = 0.0067. \quad \blacktriangleleft$$

Definition 3.6.2 has an interpretation that can be understood by considering Fig. 3.20. The joint p.d.f. f defines a surface over the xy-plane for which the height $f(x, y)$ at each point (x, y) represents the relative likelihood of that point. For instance, if it is known that $Y = y_0$, then the point (x, y) must lie on the line $y = y_0$ in the xy-plane, and the relative likelihood of any point (x, y_0) on this line is $f(x, y_0)$. Hence, the conditional p.d.f. $g_1(x|y_0)$ of X should be proportional to $f(x, y_0)$. In other words, $g_1(x|y_0)$ is essentially the same as $f(x, y_0)$, but it includes a constant factor $1/[f_2(y_0)]$, which is required to make the conditional p.d.f. integrate to unity over all values of x.

Similarly, for each value of x such that $f_1(x) > 0$, the *conditional p.d.f. of Y given that $X = x$* is defined as follows:

$$g_2(y|x) = \frac{f(x, y)}{f_1(x)} \quad \text{for } -\infty < y < \infty. \tag{3.6.5}$$

This equation is identical to Eq. (3.6.3), which was derived for discrete distributions. If $f_1(x) = 0$, then $g_2(y|x)$ is arbitrary so long as it is a p.d.f. as a function of y.

Example 3.6.6 Calculating a Conditional p.d.f. from a Joint p.d.f. Suppose that the joint p.d.f. of X and Y is as specified in Example 3.4.8 on page 122. We shall first determine the conditional p.d.f. of Y given that $X = x$ and then determine some probabilities for Y given the specific value $X = 1/2$.

The set S for which $f(x, y) > 0$ was sketched in Fig. 3.12 on page 123. Furthermore, the marginal p.d.f. f_1 was derived in Example 3.5.3 on page 132 and sketched in Fig. 3.17 on page 133. It can be seen from Fig. 3.17 that $f_1(x) > 0$ for $-1 < x < 1$ but not for $x = 0$. Therefore, for each given value of x such that $-1 < x < 0$ or $0 < x < 1$, the conditional p.d.f. $g_2(y|x)$ of Y will be as follows:

$$g_2(y|x) = \begin{cases} \dfrac{2y}{1 - x^4} & \text{for } x^2 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

In particular, if it is known that $X = 1/2$, then $\Pr\left(Y \geq \frac{1}{4} \middle| X = \frac{1}{2}\right) = 1$ and

$$\Pr\left(Y \geq \frac{3}{4} \middle| X = \frac{1}{2}\right) = \int_{3/4}^{1} g_2\left(y \middle| \frac{1}{2}\right) dy = \frac{7}{15}. \quad \blacktriangleleft$$

Note: A Conditional p.d.f. Is Not the Result of Conditioning on a Set of Probability Zero. The conditional p.d.f. $g_1(x|y)$ of X given $Y = y$ is the p.d.f. we would use for X if we were to learn that $Y = y$. This sounds as if we were conditioning on the event $\{Y = y\}$, which has zero probability if Y has a continuous distribution. Actually, for the cases we shall see in this text, the value of $g_1(x|y)$ is a limit:

$$g_1(x|y) = \lim_{\epsilon \to 0} \frac{\partial}{\partial x} \Pr(X \leq x | y - \epsilon < Y \leq y + \epsilon). \tag{3.6.6}$$

The conditioning event $\{y - \epsilon \leq Y \leq y + \epsilon\}$ in Eq. (3.6.6) has positive probability if the marginal p.d.f. of Y is positive at y. The mathematics required to make this rigorous is beyond the scope of this text. (See Exercise 11 in this section and Exercises 25 and 26 in Sec. 3.11 for results that we can prove.) Another way to think about conditioning on a continuous random variable is to notice that the conditional p.d.f.'s that we compute are typically continuous as a function of the conditioning variable. This means that conditioning on $Y = y$ or on $Y = y + \epsilon$ for small ϵ will produce nearly the same conditional distribution for X. So it does not matter much if we use $Y = y$ as a surogate for Y close to y. Nevertheless, it is important to keep in mind that the conditional p.d.f. of X given $Y = y$ is better thought of as the conditional p.d.f. of X given that Y is very close to y. This wording is awkward, so we shall not use it, but we must remember the distinction between the conditional p.d.f. and conditioning on an event with probability 0. Despite this distinction, it is still legitimate to treat Y as the constant y when dealing with the conditional distribution of X given $Y = y$.

For mixed joint distributions, we continue to use Eqs. (3.6.2) and (3.6.3) to define conditional p.f.'s and p.d.f.'s.

Definition 3.6.3

Conditional p.f. or p.d.f. from Mixed Distribution. Let X be discrete and let Y be continuous with joint p.f./p.d.f. f. Then the *conditional p.f. of X given $Y = y$* is defined by Eq. (3.6.2), and the *conditional p.d.f. of Y given $X = x$* is defined by Eq. (3.6.3).

Construction of the Joint Distribution

Example 3.6.7

Defective Parts. Suppose that a certain machine produces defective and nondefective parts, but we do not know what proportion of defectives we would find among all parts that could be produced by this machine. Let P stand for the unknown proportion of defective parts among all possible parts produced by the machine. If we were to learn that $P = p$, we might be willing to say that the parts were independent of each other and each had probability p of being defective. In other words, if we condition on $P = p$, then we have the situation described in Example 3.1.9. As in that example, suppose that we examine n parts and let X stand for the number of defectives among the n examined parts. The distribution of X, assuming that we know $P = p$, is the binomial distribution with parameters n and p. That is, we can let the binomial p.f. (3.1.4) be the conditional p.f. of X given $P = p$, namely,

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

We might also believe that P has a continuous distribution with p.d.f. such as $f_2(p) = 1$ for $0 \le p \le 1$. (This means that P has the uniform distribution on the interval $[0, 1]$.) We know that the conditional p.f. g_1 of X given $P = p$ satisfies

$$g_1(x|p) = \frac{f(x, p)}{f_2(p)},$$

where f is the joint p.f./p.d.f. of X and P. If we multiply both sides of this equation by $f_2(p)$, it follows that the joint p.f./p.d.f. of X and P is

$$f(x, p) = g_1(x|p) f_2(p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for } x = 0, \ldots, n, \text{ and } 0 \le p \le 1.$$

◀

The construction in Example 3.6.7 is available in general, as we explain next.

Generalizing the Multiplication Rule for Conditional Probabilities A special case of Theorem 2.1.2, the multiplication rule for conditional probabilities, says that if A and B are two events, then $\Pr(A \cap B) = \Pr(A) \Pr(B|A)$. The following theorem, whose proof is immediate from Eqs. (3.6.4) and (3.6.5), generalizes Theorem 2.1.2 to the case of two random variables.

Theorem 3.6.2 Multiplication Rule for Distributions. Let X and Y be random variables such that X has p.f. or p.d.f. $f_1(x)$ and Y has p.f. or p.d.f. $f_2(y)$. Also, assume that the conditional p.f. or p.d.f. of X given $Y = y$ is $g_1(x|y)$ while the conditional p.f. or p.d.f. of Y given $X = x$ is $g_2(y|x)$. Then for each y such that $f_2(y) > 0$ and each x,

$$f(x, y) = g_1(x|y) f_2(y), \tag{3.6.7}$$

where f is the joint p.f., p.d.f., or p.f./p.d.f. of X and Y. Similarly, for each x such that $f_1(x) > 0$ and each y,

$$f(x, y) = f_1(x) g_2(y|x). \tag{3.6.8}$$

■

In Theorem 3.6.2, if $f_2(y_0) = 0$ for some value y_0, then it can be assumed without loss of generality that $f(x, y_0) = 0$ for all values of x. In this case, both sides of Eq. (3.6.7) will be 0, and the fact that $g_1(x|y_0)$ is not uniquely defined becomes irrelevant. Hence, Eq. (3.6.7) will be satisfied for *all* values of x and y. A similar statement applies to Eq. (3.6.8).

Example 3.6.8 Waiting in a Queue. Let X be the amount of time that a person has to wait for service in a queue. The faster the server works in the queue, the shorter should be the waiting time. Let Y stand for the rate at which the server works, which we will take to be unknown. A common choice of conditional distribution for X given $Y = y$ has conditional p.d.f. for each $y > 0$:

$$g_1(x|y) = \begin{cases} ye^{-xy} & \text{for } x \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

We shall assume that Y has a continuous distribution with p.d.f. $f_2(y) = e^{-y}$ for $y > 0$. Now we can construct the joint p.d.f. of X and Y using Theorem 3.6.2:

$$f(x, y) = g_1(x|y) f_2(y) = \begin{cases} ye^{-y(x+1)} & \text{for } x \ge 0, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

◀

Example 3.6.9 Defective Parts. Let X be the number of defective parts in a sample of size n, and let P be the proportion of defectives among all parts, as in Example 3.6.7. The joint p.f./p.d.f of X and $P = p$ was calculated there as

$$f(x, p) = g_1(x|p) f_2(p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{for } x = 0, \ldots, n \text{ and } 0 \le p \le 1.$$

We could now compute the conditional p.d.f. of P given $X = x$ by first finding the marginal p.f. of X:

$$f_1(x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp, \tag{3.6.9}$$

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

$$g_2(p|x) = \frac{f(x, p)}{f_1(x)} = \frac{p^x (1-p)^{n-x}}{\int_0^1 q^x (1-q)^{n-x} dq}, \quad \text{for } 0 < p < 1. \tag{3.6.10}$$

The integral in the denominator of Eq. (3.6.10) can be tedious to calculate, but it can be found. For example, if $n = 2$ and $x = 1$, we get

$$\int_0^1 q(1-q) dq = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

In this case, $g_2(p|1) = 6p(1-p)$ for $0 \le p \le 1$. ◀

Bayes' Theorem and the Law of Total Probability for Random Variables The calculation done in Eq. (3.6.9) is an example of the generalization of the law of total probability to random variables. Also, the calculation in Eq. (3.6.10) is an example of the generalization of Bayes' theorem to random variables. The proofs of these results are straightforward and not given here.

Theorem 3.6.3 Law of Total Probability for Random Variables. If $f_2(y)$ is the marginal p.f. or p.d.f. of a random variable Y and $g_1(x|y)$ is the conditional p.f. or p.d.f. of X given $Y = y$, then the marginal p.f. or p.d.f. of X is

$$f_1(x) = \sum_y g_1(x|y) f_2(y), \tag{3.6.11}$$

if Y is discrete. If Y is continuous, the marginal p.f. or p.d.f. of X is

$$f_1(x) = \int_{-\infty}^{\infty} g_1(x|y) f_2(y)\, dy. \tag{3.6.12}$$

■

There are versions of Eqs. (3.6.11) and (3.6.12) with x and y switched and the subscripts 1 and 2 switched. These versions would be used if the joint distribution of X and Y were constructed from the conditional distribution of Y given X and the marginal distribution of X.

Theorem 3.6.4 Bayes' Theorem for Random Variables. If $f_2(y)$ is the marginal p.f. or p.d.f. of a random variable Y and $g_1(x|y)$ is the conditional p.f. or p.d.f. of X given $Y = y$, then the conditional p.f. or p.d.f. of Y given $X = x$ is

$$g_2(y|x) = \frac{g_1(x|y) f_2(y)}{f_1(x)}, \tag{3.6.13}$$

where $f_1(x)$ is obtained from Eq. (3.6.11) or (3.6.12). Similarly, the conditional p.f. or p.d.f. of X given $Y = y$ is

$$g_1(x|y) = \frac{g_2(y|x) f_1(x)}{f_2(y)}, \tag{3.6.14}$$

where $f_2(y)$ is obtained from Eq. (3.6.11) or (3.6.12) with x and y switched and with the subscripts 1 and 2 switched. ■

Example 3.6.10

Choosing Points from Uniform Distributions. Suppose that a point X is chosen from the uniform distribution on the interval $[0, 1]$, and that after the value $X = x$ has been observed $(0 < x < 1)$, a point Y is then chosen from the uniform distribution on the interval $[x, 1]$. We shall derive the marginal p.d.f. of Y.

Since X has a uniform distribution, the marginal p.d.f. of X is as follows:

$$f_1(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for each value $X = x$ $(0 < x < 1)$, the conditional distribution of Y is the uniform distribution on the interval $[x, 1]$. Since the length of this interval is $1 - x$, the conditional p.d.f. of Y given that $X = x$ will be

$$g_2(y|x) = \begin{cases} \dfrac{1}{1-x} & \text{for } x < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

It follows from Eq. (3.6.8) that the joint p.d.f. of X and Y will be

$$f(x, y) = \begin{cases} \dfrac{1}{1-x} & \text{for } 0 < x < y < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.6.15}$$

Thus, for $0 < y < 1$, the value of the marginal p.d.f. $f_2(y)$ of Y will be

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^y \frac{1}{1-x}\, dx = -\log(1 - y). \tag{3.6.16}$$

Furthermore, since Y cannot be outside the interval $0 < y < 1$, then $f_2(y) = 0$ for $y \le 0$ or $y \ge 1$. This marginal p.d.f. f_2 is sketched in Fig. 3.21. It is interesting to note that in this example the function f_2 is unbounded.

We can also find the conditional p.d.f. of X given $Y = y$ by applying Bayes' theorem (3.6.14). The product of $g_2(y|x)$ and $f_1(x)$ was already calculated in Eq. (3.6.15).

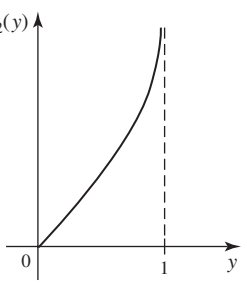

Figure 3.21 The marginal p.d.f. of Y in Example 3.6.10.

The ratio of this product to $f_2(y)$ from Eq. (3.6.16) is

$$g_1(x|y) = \begin{cases} \dfrac{-1}{(1-x)\log(1-y)} & \text{for } 0 < x < y, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Theorem 3.6.5 Independent Random Variables. Suppose that X and Y are two random variables having a joint p.f., p.d.f., or p.f./p.d.f. f. Then X and Y are independent if and only if for every value of y such that $f_2(y) > 0$ and every value of x,

$$g_1(x|y) = f_1(x). \tag{3.6.17}$$

Proof Theorem 3.5.4 says that X and Y are independent if and only if $f(x, y)$ can be factored in the following form for $-\infty < x < \infty$ and $-\infty < y < \infty$:

$$f(x, y) = f_1(x) f_2(y),$$

which holds if and only if, for all x and all y such that $f_2(y) > 0$,

$$f_1(x) = \frac{f(x, y)}{f_2(y)}. \tag{3.6.18}$$

But the right side of Eq. (3.6.18) is the formula for $g_1(x|y)$. Hence, X and Y are independent if and only if Eq. (3.6.17) holds for all x and all y such that $f_2(y) > 0$. ■

Theorem 3.6.5 says that X and Y are independent if and only if the conditional p.f. or p.d.f. of X given $Y = y$ is the same as the marginal p.f. or p.d.f. of X for all y such that $f_2(y) > 0$. Because $g_1(x|y)$ is arbitrary when $f_2(y) = 0$, we cannot expect Eq. (3.6.17) to hold in that case.

Similarly, it follows from Eq. (3.6.8) that X and Y are independent if and only if

$$g_2(y|x) = f_2(y), \tag{3.6.19}$$

for every value of x such that $f_1(x) > 0$. Theorem 3.6.5 and Eq. (3.6.19) give the mathematical justification for the meaning of independence that we presented on page 136.

Note: Conditional Distributions Behave Just Like Distributions. As we noted on page 59, conditional probabilities behave just like probabilities. Since distributions are just collections of probabilities, it follows that conditional distributions behave just like distributions. For example, to compute the conditional probability that a discrete random variable X is in some interval $[a, b]$ given $Y = y$, we must add $g_1(x|y)$ for all values of x in the interval. Also, theorems that we have proven or shall prove about distributions will have versions conditional on additional random variables. We shall postpone examples of such theorems until Sec. 3.7 because they rely on joint distributions of more than two random variables.

Summary

The conditional distribution of one random variable X given an observed value y of another random variable Y is the distribution we would use for X if we were to learn that $Y = y$. When dealing with the conditional distribution of X given $Y = y$, it is safe to behave as if Y were the constant y. If X and Y have joint p.f., p.d.f., or p.f./p.d.f. $f(x, y)$, then the conditional p.f. or p.d.f. of X given $Y = y$ is $g_1(x|y) =$

$f(x, y)/f_2(y)$, where f_2 is the marginal p.f. or p.d.f. of Y. When it is convenient to specify a conditional distribution directly, the joint distribution can be constructed from the conditional together with the other marginal. For example,

$$f(x, y) = g_1(x|y)f_2(y) = f_1(x)g_2(y|x).$$

In this case, we have versions of the law of total probability and Bayes' theorem for random variables that allow us to calculate the other marginal and conditional.

Two random variables X and Y are independent if and only if the conditional p.f. or p.d.f. of X given $Y = y$ is the same as the marginal p.f. or p.d.f. of X for all y such that $f_2(y) > 0$. Equivalently, X and Y are independent if and only if the conditional p.f. of p.d.f. of Y given $X = x$ is the same as the marginal p.f. or p.d.f. of Y for all x such that $f_1(x) > 0$.

Exercises

1. Suppose that two random variables X and Y have the joint p.d.f. in Example 3.5.10 on page 139. Compute the conditional p.d.f. of X given $Y = y$ for each y.

2. Each student in a certain high school was classified according to her year in school (freshman, sophomore, junior, or senior) and according to the number of times that she had visited a certain museum (never, once, or more than once). The proportions of students in the various classifications are given in the following table:

	Never	Once	More than once
Freshmen	0.08	0.10	0.04
Sophomores	0.04	0.10	0.04
Juniors	0.04	0.20	0.09
Seniors	0.02	0.15	0.10

a. If a student selected at random from the high school is a junior, what is the probability that she has never visited the museum?

b. If a student selected at random from the high school has visited the museum three times, what is the probability that she is a senior?

3. Suppose that a point (X, Y) is chosen at random from the disk S defined as follows:

$$S = \{(x, y) : (x - 1)^2 + (y + 2)^2 \le 9\}.$$

Determine **(a)** the conditional p.d.f. of Y for every given value of X, and **(b)** $\Pr(Y > 0|X = 2)$.

4. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} c(x + y^2) & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the conditional p.d.f. of X for every given value of Y, and **(b)** $\Pr(X < \frac{1}{2}|Y = \frac{1}{2})$.

5. Suppose that the joint p.d.f. of two points X and Y chosen by the process described in Example 3.6.10 is as given by Eq. (3.6.15). Determine **(a)** the conditional p.d.f. of X for every given value of Y, and **(b)** $\Pr\left(X > \frac{1}{2}\middle|Y = \frac{3}{4}\right)$.

6. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} c \sin x & \text{for } 0 \le x \le \pi/2 \text{ and } 0 \le y \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the conditional p.d.f. of Y for every given value of X, and **(b)** $\Pr(1 < Y < 2|X = 0.73)$.

7. Suppose that the joint p.d.f. of two random variables X and Y is as follows:

$$f(x, y) = \begin{cases} \frac{3}{16}(4 - 2x - y) & \text{for } x > 0,\ y > 0, \\ & \quad \text{and } 2x + y < 4, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the conditional p.d.f. of Y for every given value of X, and **(b)** $\Pr(Y \ge 2|X = 0.5)$.

8. Suppose that a person's score X on a mathematics aptitude test is a number between 0 and 1, and that his score Y on a music aptitude test is also a number between 0 and 1. Suppose further that in the population of all college students in the United States, the scores X and Y are distributed according to the following joint p.d.f.:

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y) & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

a. What proportion of college students obtain a score greater than 0.8 on the mathematics test?

b. If a student's score on the music test is 0.3, what is the probability that his score on the mathematics test will be greater than 0.8?

c. If a student's score on the mathematics test is 0.3, what is the probability that his score on the music test will be greater than 0.8?

9. Suppose that either of two instruments might be used for making a certain measurement. Instrument 1 yields a measurement whose p.d.f. h_1 is

$$h_1(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Instrument 2 yields a measurement whose p.d.f. h_2 is

$$h_2(x) = \begin{cases} 3x^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that one of the two instruments is chosen at random and a measurement X is made with it.

a. Determine the marginal p.d.f. of X.

b. If the value of the measurement is $X = 1/4$, what is the probability that instrument 1 was used?

10. In a large collection of coins, the probability X that a head will be obtained when a coin is tossed varies from one coin to another, and the distribution of X in the collection is specified by the following p.d.f.:

$$f_1(x) = \begin{cases} 6x(1-x) & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a coin is selected at random from the collection and tossed once, and that a head is obtained. Determine the conditional p.d.f. of X for this coin.

11. The definition of the conditional p.d.f. of X given $Y = y$ is arbitrary if $f_2(y) = 0$. The reason that this causes no serious problem is that it is highly unlikely that we will observe Y close to a value y_0 such that $f_2(y_0) = 0$. To be more precise, let $f_2(y_0) = 0$, and let $A_0 = [y_0 - \epsilon, y_0 + \epsilon]$. Also, let y_1 be such that $f_2(y_1) > 0$, and let $A_1 = [y_1 - \epsilon, y_1 + \epsilon]$. Assume that f_2 is continuous at both y_0 and y_1. Show that

$$\lim_{\epsilon \to 0} \frac{\Pr(Y \in A_0)}{\Pr(Y \in A_1)} = 0.$$

That is, the probability that Y is close to y_0 is much smaller than the probability that Y is close to y_1.

12. Let Y be the rate (calls per hour) at which calls arrive at a switchboard. Let X be the number of calls during a two-hour period. Suppose that the marginal p.d.f. of Y is

$$f_2(y) = \begin{cases} e^{-y} & \text{if } y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and that the conditional p.f. of X given $Y = y$ is

$$g_1(x|y) = \begin{cases} \dfrac{(2y)^x}{x!} e^{-2y} & \text{if } x = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

a. Find the marginal p.f. of X. (You may use the formula $\int_0^\infty y^k e^{-y}\, dy = k!$.)

b. Find the conditional p.d.f. $g_2(y|0)$ of Y given $X = 0$.

c. Find the conditional p.d.f. $g_2(y|1)$ of Y given $X = 1$.

d. For what values of y is $g_2(y|1) > g_2(y|0)$? Does this agree with the intuition that the more calls you see, the higher you should think the rate is?

13. Start with the joint distribution of treatment group and response in Table 3.6 on page 138. For each treatment group, compute the conditional distribution of response given the treatment group. Do they appear to be very similar or quite different?

3.7 Multivariate Distributions

In this section, we shall extend the results that were developed in Sections 3.4, 3.5, and 3.6 for two random variables X and Y to an arbitrary finite number n of random variables $X_1, \ldots, X_n$. In general, the joint distribution of more than two random variables is called a multivariate distribution. *The theory of statistical inference (the subject of the part of this book beginning with Chapter 7) relies on mathematical models for observable data in which each observation is a random variable. For this reason, multivariate distributions arise naturally in the mathematical models for data. The most commonly used model will be one in which the individual data random variables are conditionally independent given one or two other random variables.*

Joint Distributions

Example 3.7.1

A Clinical Trial. Suppose that m patients with a certain medical condition are given a treatment, and each patient either recovers from the condition or fails to recover. For each $i = 1, \ldots, m$, we can let $X_i = 1$ if patient i recovers and $X_i = 0$ if not. We might also believe that there is a random variable P having a continuous distribution taking values between 0 and 1 such that, if we knew that $P = p$, we would say that the m patients recover or fail to recover independently of each other each with probability p of recovery. We now have named $n = m + 1$ random variables in which we are interested. ◀

The situation described in Example 3.7.1 requires us to construct a joint distribution for n random variables. We shall now provide definitions and examples of the important concepts needed to discuss multivariate distributions.

Definition 3.7.1

Joint Distribution Function/c.d.f. The *joint c.d.f.* of n random variables $X_1, \ldots, X_n$ is the function F whose value at every point $(x_1, \ldots, x_n)$ in n-dimensional space R^n is specified by the relation

$$F(x_1, \ldots, x_n) = \Pr(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n). \tag{3.7.1}$$

Every multivariate c.d.f. satisfies properties similar to those given earlier for univariate and bivariate c.d.f.'s.

Example 3.7.2

Failure Times. Suppose that a machine has three parts, and part i will fail at time X_i for $i = 1, 2, 3$. The following function might be the joint c.d.f. of X_1, X_2, and X_3:

$$F(x_1, x_2, x_3) = \begin{cases} (1 - e^{-x_1})(1 - e^{-2x_2})(1 - e^{-3x_3}) & \text{for } x_1, x_2, x_3 \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Vector Notation In the study of the joint distribution of n random variables $X_1, \ldots, X_n$, it is often convenient to use the vector notation $\boldsymbol{X} = (X_1, \ldots, X_n)$ and to refer to $\boldsymbol{X}$ as a *random vector*. Instead of speaking of the joint distribution of the random variables $X_1, \ldots, X_n$ with a joint c.d.f. $F(x_1, \ldots, x_n)$, we can simply speak of the distribution of the random vector $\boldsymbol{X}$ with c.d.f. $F(\boldsymbol{x})$. When this vector notation is used, it must be kept in mind that if $\boldsymbol{X}$ is an n-dimensional random vector, then its c.d.f. is defined as a function on n-dimensional space R^n. At each point $\boldsymbol{x} = (x_1, \ldots, x_n) \in R^n$, the value of $F(\boldsymbol{x})$ is specified by Eq. (3.7.1).

Definition 3.7.2

Joint Discrete Distribution/p.f. It is said that n random variables $X_1, \ldots, X_n$ have a *discrete joint distribution* if the random vector $(X_1, \ldots, X_n)$ can have only a finite number or an infinite sequence of different possible values $(x_1, \ldots, x_n)$ in R^n. The *joint p.f.* of $X_1, \ldots, X_n$ is then defined as the function f such that for every point $(x_1, \ldots, x_n) \in R^n$,

$$f(x_1, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n).$$

In vector notation, Definition 3.7.2 says that the random vector $\boldsymbol{X}$ has a discrete distribution and that its p.f. is specified at every point $\boldsymbol{x} \in R^n$ by the relation

$$f(\boldsymbol{x}) = \Pr(\boldsymbol{X} = \boldsymbol{x}).$$

The following result is a simple generalization of Theorem 3.4.2.

Theorem 3.7.1 If $\boldsymbol{X}$ has a joint discrete distribution with joint p.f. f, then for every subset $C \subset R^n$,

$$\Pr(\boldsymbol{X} \in C) = \sum_{x \in C} f(\boldsymbol{x}).$$ ■

It is easy to show that, if each of $X_1, \ldots, X_n$ has a discrete distribution, then $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a discrete joint distribution.

Example 3.7.3 A Clinical Trial. Consider the m patients in Example 3.7.1. Suppose for now that $P = p$ is known so that we don't treat it as a random variable. The joint p.f. of $\boldsymbol{X} = (X_1, \ldots, X_m)$ is

$$f(\boldsymbol{x}) = p^{x_1 + \cdots + x_m}(1-p)^{m - x_1 - \cdots - x_m},$$

for all $x_i \in \{0, 1\}$ and 0 otherwise. ◀

Definition 3.7.3 Continuous Distribution/p.d.f. It is said that n random variables $X_1, \ldots, X_n$ have a continuous joint distribution if there is a nonnegative function f defined on R^n such that for every subset $C \subset R^n$,

$$\Pr[(X_1, \ldots, X_n) \in C] = \int \cdots \int_C f(x_1, \ldots x_n)\, dx_1 \cdots dx_n, \tag{3.7.2}$$

if the integral exists. The function f is called the *joint p.d.f.* of $X_1, \ldots, X_n$.

In vector notation, $f(\boldsymbol{x})$ denotes the p.d.f. of the random vector $\boldsymbol{X}$ and Eq. (3.7.2) could be rewritten more simply in the form

$$\Pr(\boldsymbol{X} \in C) = \int \cdots \int_C f(\boldsymbol{x})\, d\boldsymbol{x}.$$

Theorem 3.7.2 If the joint distribution of $X_1, \ldots, X_n$ is continuous, then the joint p.d.f. f can be derived from the joint c.d.f. F by using the relation

$$f(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}$$

at all points $(x_1, \ldots, x_n)$ at which the derivative in this relation exists. ■

Example 3.7.4 Failure Times. We can find the joint p.d.f. for the three random variables in Example 3.7.2 by applying Theorem 3.7.2. The third-order mixed partial is easily calculated to be

$$f(x_1, x_2, x_3) = \begin{cases} 6e^{-x_1 - 2x_2 - 3x_3} & \text{for } x_1, x_2, x_3 > 0, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

It is important to note that, even if each of $X_1, \ldots, X_n$ has a continuous distribution, the vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ might not have a continuous joint distribution. See Exercise 9 in this section.

Example 3.7.5 Service Times in a Queue. A queue is a system in which customers line up for service and receive their service according to some algorithm. A simple model is the single-server queue, in which all customers wait for a single server to serve everyone ahead of them in the line and then they get served. Suppose that n customers arrive at a

single-server queue for service. Let X_i be the time that the server spends serving customer i for $i = 1, \ldots, n$. We might use a joint distribution for $\boldsymbol{X} = (X_1, \ldots, X_n)$ with joint p.d.f. of the form

$$f(\boldsymbol{x}) = \begin{cases} \dfrac{c}{\left(2 + \sum_{i=1}^{n} x_i\right)^{n+1}} & \text{for all } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.3}$$

We shall now find the value of c such that the function in Eq. (3.7.3) is a joint p.d.f. We can do this by integrating over each variable $x_1, \ldots, x_n$ in succession (starting with x_n). The first integral is

$$\int_0^\infty \frac{c}{(2 + x_1 + \cdots + x_n)^{n+1}}\, dx_n = \frac{c/n}{(2 + x_1 + \cdots + x_{n-1})^n}. \tag{3.7.4}$$

The right-hand side of Eq. (3.7.4) is in the same form as the original p.d.f. except that n has been reduced to $n - 1$ and c has been divided by n. It follows that when we integrate over the variable x_i (for $i = n - 1, n - 2, \ldots, 1$), the result will be in the same form with n reduced to $i - 1$ and c divided by $n(n - 1) \cdots i$. The result of integrating all coordinates except x_1 is then

$$\frac{c/n!}{(2 + x_1)^2}, \quad \text{for } x_1 > 0.$$

Integrating x_1 out of this yields $c/[2(n!)]$, which must equal 1, so $c = 2(n!)$. ◀

Mixed Distributions

Example 3.7.6 Arrivals at a Queue. In Example 3.7.5, we introduced the single-server queue and discussed service times. Some features that influence the performance of a queue are the rate at which customers arrive and the rate at which customers are served. Let Z stand for the rate at which customers are served, and let Y stand for the rate at which customers arrive at the queue. Finally, let W stand for the number of customers that arrive during one day. Then W is discrete while Y and Z could be continuous random variables. A possible joint p.f./p.d.f. for these three random variables is

$$f(y, z, w) = \begin{cases} 6e^{-3z-10y}(8y)^w/w! & \text{for } z, y > 0 \text{ and } w = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

We can verify this claim shortly. ◀

Definition 3.7.4 Joint p.f./p.d.f. Let $X_1, \ldots, X_n$ be random variables, some of which have a continuous joint distribution and some of which have discrete distributions; their joint distribution would then be represented by a function f that we call the *joint p.f./p.d.f.*. The function has the property that the probability that $\boldsymbol{X}$ lies in a subset $C \subset R^n$ is calculated by summing $f(\boldsymbol{x})$ over the values of the coordinates of $\boldsymbol{x}$ that correspond to the discrete random variables and integrating over those coordinates that correspond to the continuous random variables for all points $\boldsymbol{x} \in C$.

Example 3.7.7 Arrivals at a Queue. We shall now verify that the proposed p.f./p.d.f. in Example 3.7.6 actually sums and integrates to 1 over all values of (y, z, w). We must sum over w and integrate over y and z. We have our choice of in what order to do them. It is not

difficult to see that we can factor f as $f(y, z, w) = h_2(z)h_{13}(y, w)$, where

$$h_2(z) = \begin{cases} 6e^{-3z} & \text{for } z > 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$h_{13}(y, w) = \begin{cases} e^{-10y}(8y)^w/w! & \text{for } y > 0 \text{ and } w = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

So we can integrate z out first to get

$$\int_{-\infty}^{\infty} f(y, z, w)dz = h_{13}(y, w) \int_0^{\infty} 6e^{-3z}dz = 2h_{13}(y, w).$$

Integrating y out of $h_{13}(y, w)$ is possible, but not pleasant. Instead, notice that $(8y)^w/w!$ is the wth term in the Taylor expansion of e^{8y}. Hence,

$$\sum_{w=0}^{\infty} 2h_{13}(y, w) = 2e^{-10y} \sum_{w=0}^{\infty} \frac{(8y)^w}{w!} = 2e^{-10y}e^{8y} = 2e^{-2y},$$

for $y > 0$ and 0 otherwise. Finally, integrating over y yields 1. ◀

Example 3.7.8

A Clinical Trial. In Example 3.7.1, one of the random variables P has a continuous distribution, and the others $X_1, \ldots, X_m$ have discrete distributions. A possible joint p.f./p.d.f. for $(X_1, \ldots, X_m, P)$ is

$$f(\boldsymbol{x}, p) = \begin{cases} p^{x_1+\cdots+x_m}(1-p)^{m-x_1-\cdots-x_m} & \text{for all } x_i \in \{0, 1\} \text{ and } 0 \le p \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We can find probabilities based on this function. Suppose, for example, that we want the probability that there is exactly one success among the first two patients, that is, $\Pr(X_1 + X_2 = 1)$. We must integrate $f(\boldsymbol{x}, p)$ over p and sum over all values of $\boldsymbol{x}$ that have $x_1 + x_2 = 1$. For purposes of illustration, suppose that $m = 4$. First, factor out $p^{x_1+x_2}(1-p)^{2-x_1-x_2} = p(1-p)$, which yields

$$f(\boldsymbol{x}, p) = [p(1-p)]p^{x_3+x_4}(1-p)^{2-x_3-x_4},$$

for $x_3, x_4 \in \{0, 1\}$, $0 < p < 1$, and $x_1 + x_2 = 1$. Summing over x_3 yields

$$[p(1-p)]\left(p^{x_4}(1-p)^{1-x_4}(1-p) + pp^{x_4}(1-p)^{1-x_4}\right) = [p(1-p)]p^{x_4}(1-p)^{1-x_4}.$$

Summing this over x_4 gives $p(1-p)$. Next, integrate over p to get $\int_0^1 p(1-p)dp = 1/6$. Finally, note that there are two (x_1, x_2) vectors, $(1, 0)$ and $(0, 1)$, that have $x_1 + x_2 = 1$, so $\Pr(X_1 + X_2 = 1) = (1/6) + (1/6) = 1/3$. ◀

Marginal Distributions

Deriving a Marginal p.d.f. If the joint distribution of n random variables $X_1, \ldots, X_n$ is known, then the marginal distribution of each single random variable X_i can be derived from this joint distribution. For example, if the joint p.d.f. of $X_1, \ldots, X_n$ is f, then the marginal p.d.f. f_1 of X_1 is specified at every value x_1 by the relation

$$f_1(x_1) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1} f(x_1, \ldots, x_n)\, dx_2 \cdots dx_n.$$

More generally, the marginal joint p.d.f. of any k of the n random variables $X_1, \ldots, X_n$ can be found by integrating the joint p.d.f. over all possible values of

the other $n - k$ variables. For example, if f is the joint p.d.f. of four random variables X_1, X_2, X_3, and X_4, then the marginal bivariate p.d.f. f_{24} of X_2 and X_4 is specified at each point (x_2, x_4) by the relation

$$f_{24}(x_2, x_4) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_1\, dx_3.$$

Example 3.7.9

Service Times in a Queue. Suppose that $n = 5$ in Example 3.7.5 and that we want the marginal bivariate p.d.f. of (X_1, X_4). We must integrate Eq. (3.7.3) over x_2, x_3, and x_5. Since the joint p.d.f. is symmetric with respect to permutations of the coordinates of $\boldsymbol{x}$, we shall just integrate over the last three variables and then change the names of the remaining variables to x_1 and x_4. We already saw how to do this in Example 3.7.5. The result is

$$f_{12}(x_1, x_2) = \begin{cases} \dfrac{4}{(2 + x_1 + x_2)^3} & \text{for } x_1, x_2 > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.5}$$

Then f_{14} is just like (3.7.5) with all the 2 subscripts changed to 4. The univariate marginal p.d.f. of each X_i is

$$f_i(x_i) = \begin{cases} \dfrac{2}{(2 + x_i)^2} & \text{for } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.6}$$

So, for example, if we want to know how likely it is that a customer will have to wait longer than three time units, we can calculate $\Pr(X_i > 3)$ by integrating the function in Eq. (3.7.6) from 3 to ∞. The result is 0.4. ◀

If n random variables $X_1, \ldots, X_n$ have a discrete joint distribution, then the marginal joint p.f. of each subset of the n variables can be obtained from relations similar to those for continuous distributions. In the new relations, the integrals are replaced by sums.

Deriving a Marginal c.d.f. Consider now a joint distribution for which the joint c.d.f. of $X_1, \ldots, X_n$ is F. The marginal c.d.f. F_1 of X_1 can be obtained from the following relation:

$$\begin{aligned} F_1(x_1) &= \Pr(X_1 \le x_1) = \Pr(X_1 \le x_1, X_2 < \infty, \ldots, X_n < \infty) \\ &= \lim_{x_2, \ldots, x_n \to \infty} F(x_1, x_2, \ldots, x_n). \end{aligned}$$

Example 3.7.10

Failure Times. We can find the marginal c.d.f. of X_1 from the joint c.d.f. in Example 3.7.2 by letting x_2 and x_3 go to ∞. The limit is $F_1(x_1) = 1 - e^{-x_1}$ for $x_1 \ge 0$ and 0 otherwise. ◀

More generally, the marginal joint c.d.f. of any k of the n random variables $X_1, \ldots, X_n$ can be found by computing the limiting value of the n-dimensional c.d.f. F as $x_j \to \infty$ for each of the other $n - k$ variables x_j. For example, if F is the joint c.d.f. of four random variables X_1, X_2, X_3, and X_4, then the marginal bivariate c.d.f. F_{24} of X_2 and X_4 is specified at every point (x_2, x_4) by the relation

$$F_{24}(x_2, x_4) = \lim_{x_1, x_3 \to \infty} F(x_1, x_2, x_3, x_4).$$

Example 3.7.11 Failure Times. We can find the marginal bivariate c.d.f. of X_1 and X_3 from the joint c.d.f. in Example 3.7.2 by letting x_2 go to ∞. The limit is

$$F_{13}(x_1, x_3) = \begin{cases} (1 - e^{-x_1})(1 - e^{-3x_3}) & \text{for } x_1, x_3 \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Independent Random Variables

Definition 3.7.5 Independent Random Variables. It is said that n random variables $X_1, \ldots, X_n$ are *independent* if, for every n sets $A_1, A_2, \ldots, A_n$ of real numbers,

$$\Pr(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \Pr(X_1 \in A_1) \Pr(X_2 \in A_2) \cdots \Pr(X_n \in A_n).$$

If $X_1, \ldots, X_n$ are independent, it follows easily that the random variables in every nonempty subset of $X_1, \ldots, X_n$ are also independent. (See Exercise 11.)

There is a generalization of Theorem 3.5.4.

Theorem 3.7.3 Let F denote the joint c.d.f. of $X_1, \ldots, X_n$, and let F_i denote the marginal univariate c.d.f. of X_i for $i = 1, \ldots, n$. The variables $X_1, \ldots, X_n$ are independent if and only if, for all points $(x_1, x_2, \ldots, x_n) \in R^n$,

$$F(x_1, x_2, \ldots, x_n) = F_1(x_1) F_2(x_2) \cdots F_n(x_n).$$ ■

Theorem 3.7.3 says that $X_1, \ldots, X_n$ are independent if and only if their joint c.d.f. is the product of their n individual marginal c.d.f.'s. It is easy to check that the three random variables in Example 3.7.2 are independent using Theorem 3.7.3.

There is also a generalization of Corollary 3.5.1.

Theorem 3.7.4 If $X_1, \ldots, X_n$ have a continuous, discrete, or mixed joint distribution for which the joint p.d.f., joint p.f., or joint p.f./p.d.f. is f, and if f_i is the marginal univariate p.d.f. or p.f. of X_i $(i = 1, \ldots, n)$, then $X_1, \ldots, X_n$ are independent if and only if the following relation is satisfied at all points $(x_1, x_2, \ldots, x_n) \in R^n$:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n). \tag{3.7.7}$$ ■

Example 3.7.12 Service Times in a Queue. In Example 3.7.9, we can multiply together the two univariate marginal p.d.f.'s of X_1 and X_2 calculated using Eq. (3.7.6) and see that the product does *not* equal the bivariate marginal p.d.f. of (X_1, X_2) in Eq. (3.7.5). So X_1 and X_2 are not independent. ◀

Definition 3.7.6 Random Samples/i.i.d./Sample Size. Consider a given probability distribution on the real line that can be represented by either a p.f. or a p.d.f. f. It is said that n random variables $X_1, \ldots, X_n$ form a *random sample* from this distribution if these random variables are independent and the marginal p.f. or p.d.f. of each of them is f. Such random variables are also said to be *independent and identically distributed*, abbreviated *i.i.d.* We refer to the number n of random variables as the *sample size*.

Definition 3.7.6 says that $X_1, \ldots, X_n$ form a random sample from the distribution represented by f if their joint p.f. or p.d.f. g is specified as follows at all points $(x_1, x_2, \ldots, x_n) \in R^n$:

$$g(x_1, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

Clearly, an i.i.d. sample cannot have a mixed joint distribution.

Example 3.7.13 Lifetimes of Light Bulbs. Suppose that the lifetime of each light bulb produced in a certain factory is distributed according to the following p.d.f.:

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

We shall determine the joint p.d.f. of the lifetimes of a random sample of n light bulbs drawn from the factory's production.

The lifetimes $X_1, \ldots, X_n$ of the selected bulbs will form a random sample from the p.d.f. f. For typographical simplicity, we shall use the notation $\exp(v)$ to denote the exponential e^v when the expression for v is complicated. Then the joint p.d.f. g of $X_1, \ldots, X_n$ will be as follows: If $x_i > 0$ for $i = 1, \ldots, n$,

$$\begin{aligned} g(x_1, \ldots, x_n) &= \prod_{i=1}^{n} f(x_i) \\ &= \left(\prod_{i=1}^{n} x_i\right) \exp\left(-\sum_{i=1}^{n} x_i\right). \end{aligned}$$

Otherwise, $g(x_1, \ldots, x_n) = 0$.

Every probability involving the n lifetimes $X_1, \ldots, X_n$ can in principle be determined by integrating this joint p.d.f. over the appropriate subset of R^n. For example, if C is the subset of points $(x_1, \ldots, x_n)$ such that $x_i > 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} x_i < a$, where a is a given positive number, then

$$\Pr\left(\sum_{i=1}^{n} X_i < a\right) = \int \cdots \int_C \left(\prod_{i=1}^{n} x_i\right) \exp\left(-\sum_{i=1}^{n} x_i\right) dx_1 \cdots dx_n. \quad \blacktriangleleft$$

The evaluation of the integral given at the end of Example 3.7.13 may require a considerable amount of time without the aid of tables or a computer. Certain other probabilities, however, can be evaluated easily from the basic properties of continuous distributions and random samples. For example, suppose that for the conditions of Example 3.7.13 it is desired to find $\Pr(X_1 < X_2 < \cdots < X_n)$. Since the random variables $X_1, \ldots, X_n$ have a continuous joint distribution, the probability that at least two of these random variables will have the same value is 0. In fact, the probability is 0 that the vector $(X_1, \ldots, X_n)$ will belong to each specific subset of R^n for which the n-dimensional volume is 0. Furthermore, since $X_1, \ldots, X_n$ are independent and identically distributed, each of these variables is equally likely to be the smallest of the n lifetimes, and each is equally likely to be the largest. More generally, if the lifetimes $X_1, \ldots, X_n$ are arranged in order from the smallest to the largest, each particular ordering of $X_1, \ldots, X_n$ is as likely to be obtained as any other ordering. Since there are $n!$ different possible orderings, the probability that the particular ordering $X_1 < X_2 < \cdots < X_n$ will be obtained is $1/n!$. Hence,

$$\Pr(X_1 < X_2 < \cdots < X_n) = \frac{1}{n!}.$$

Conditional Distributions

Suppose that n random variables $X_1, \ldots, X_n$ have a continuous joint distribution for which the joint p.d.f. is f and that f_0 denotes the marginal joint p.d.f. of the $k < n$ random variables $X_1, \ldots, X_k$. Then for all values of $x_1, \ldots, x_k$ such that $f_0(x_1, \ldots, x_k) > 0$, the conditional p.d.f. of $(X_{k+1}, \ldots, X_n)$ given that $X_1 = x_1, \ldots, X_k = x_k$ is defined

as follows:

$$g_{k+1\ldots n}(x_{k+1}, \ldots, x_n|x_1, \ldots, x_k) = \frac{f(x_1, x_2, \ldots, x_n)}{f_0(x_1, \ldots, x_k)}.$$

The definition above generalizes to arbitrary joint distributions as follows.

Definition 3.7.7

Conditional p.f., p.d.f., or p.f./p.d.f. Suppose that the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ is divided into two subvectors $\boldsymbol{Y}$ and $\boldsymbol{Z}$, where $\boldsymbol{Y}$ is a k-dimensional random vector comprising k of the n random variables in $\boldsymbol{X}$, and $\boldsymbol{Z}$ is an $(n-k)$-dimensional random vector comprising the other $n-k$ random variables in $\boldsymbol{X}$. Suppose also that the n-dimensional joint p.f., p.d.f., or p.f./p.d.f. of $(\boldsymbol{Y}, \boldsymbol{Z})$ is f and that the marginal $(n-k)$-dimensional p.f., p.d.f., or p.f./p.d.f. of $\boldsymbol{Z}$ is f_2. Then for every given point $\boldsymbol{z} \in R^{n-k}$ such that $f_2(\boldsymbol{z}) > 0$, the conditional k-dimensional p.f., p.d.f., or p.f./p.d.f. g_1 of $\boldsymbol{Y}$ given $\boldsymbol{Z} = \boldsymbol{z}$ is defined as follows:

$$g_1(\boldsymbol{y}|\boldsymbol{z}) = \frac{f(\boldsymbol{y}, \boldsymbol{z})}{f_2(\boldsymbol{z})} \qquad \text{for } \boldsymbol{y} \in R^k. \tag{3.7.8}$$

Eq. (3.7.8) can be rewritten as

$$f(\boldsymbol{y}, \boldsymbol{z}) = g_1(\boldsymbol{y}|\boldsymbol{z}) f_2(\boldsymbol{z}), \tag{3.7.9}$$

which allows construction of the joint distribution from a conditional distribution and a marginal distribution. As in the bivariate case, it is safe to assume that $f(\boldsymbol{y}, \boldsymbol{z}) = 0$ whenever $f_2(\boldsymbol{z}) = 0$. Then Eq. (3.7.9) holds for all $\boldsymbol{y}$ and $\boldsymbol{z}$ even though $g_1(\boldsymbol{y}|\boldsymbol{z})$ is not uniquely defined.

Example 3.7.14

Service Times in a Queue. In Example 3.7.9, we calculated the marginal bivariate distribution of two service times $\boldsymbol{Z} = (X_1, X_2)$. We can now find the conditional three-dimensional p.d.f. of $\boldsymbol{Y} = (X_3, X_4, X_5)$ given $\boldsymbol{Z} = (x_1, x_2)$ for every pair (x_1, x_2) such that $x_1, x_2 > 0$:

$$\begin{aligned} g_1(x_3, x_4, x_5|x_1, x_2) &= \frac{f(x_1, \ldots, x_5)}{f_{12}(x_1, x_2)} \\ &= \left(\frac{240}{(2 + x_1 + \cdots + x_5)^6}\right)\left(\frac{4}{(2 + x_1 + x_2)^3}\right)^{-1} \\ &= \frac{60(2 + x_1 + x_2)^3}{(2 + x_1 + \cdots + x_5)^6}, \end{aligned} \tag{3.7.10}$$

for $x_3, x_4, x_5 > 0$, and 0 otherwise. The joint p.d.f. in (3.7.10) looks like a bunch of symbols, but it can be quite useful. Suppose that we observe $X_1 = 4$ and $X_2 = 6$. Then

$$g_1(x_3, x_4, x_5|4.6) = \begin{cases} \dfrac{103{,}680}{(12 + x_3 + x_4 + x_5)^6} & \text{for } x_3, x_4, x_5 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We can now calculate the conditional probability that $X_3 > 3$ given $X_1 = 4$, $X_2 = 6$:

$$\Pr(X_3 > 3|X_1 = 4,\ X_2 = 6) = \int_3^\infty \int_0^\infty \int_0^\infty \frac{10{,}360}{(12 + x_3 + x_4 + x_5)^6} dx_5 dx_4 dx_3$$
$$= \int_3^\infty \int_0^\infty \frac{20{,}736}{(12 + x_3 + x_4)^5} dx_4 dx_3$$
$$= \int_3^\infty \frac{5184}{(12 + x_3)^4} dx_3$$
$$= \frac{1728}{15^3} = 0.512.$$

Compare this to the calculation of $\Pr(X_3 > 3) = 0.4$ at the end of Example 3.7.9. After learning that the first two service times are a bit longer than three time units, we revise the probability that $X_3 > 3$ upward to reflect what we learned from the first two observations. If the first two service times had been small, the conditional probability that $X_3 > 3$ would have been smaller than 0.4. For example, $\Pr(X_3 > 3|X_1 = 1,\ X_2 = 1.5) = 0.216$. ◀

Example 3.7.15 Determining a Marginal Bivariate p.d.f. Suppose that Z is a random variable for which the p.d.f. f_0 is as follows:

$$f_0(z) = \begin{cases} 2e^{-2z} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.11}$$

Suppose, furthermore, that for every given value $Z = z > 0$ two other random variables X_1 and X_2 are independent and identically distributed and the conditional p.d.f. of each of these variables is as follows:

$$g(x|z) = \begin{cases} ze^{-zx} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.12}$$

We shall determine the marginal joint p.d.f. of (X_1, X_2).

Since X_1 and X_2 are i.i.d. for each given value of Z, their conditional joint p.d.f. when $Z = z > 0$ is

$$g_{12}(x_1, x_2|z) = \begin{cases} z^2 e^{-z(x_1+x_2)} & \text{for } x_1, x_2 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The joint p.d.f. f of (Z, X_1, X_2) will be positive only at those points (z, x_1, x_2) such that $x_1, x_2, z > 0$. It now follows that, at every such point,

$$f(z, x_1, x_2) = f_0(z) g_{12}(x_1, x_2|z) = 2z^2 e^{-z(2+x_1+x_2)}.$$

For $x_1 > 0$ and $x_2 > 0$, the marginal joint p.d.f. $f_{12}(x_1, x_2)$ of X_1 and X_2 can be determined either using integration by parts or some special results that will arise in Sec. 5.7:

$$f_{12}(x_1, x_2) = \int_0^\infty f(z, x_1, x_2)\, dz = \frac{4}{(2 + x_1 + x_2)^3},$$

for $x_1, x_2 > 0$. The reader will note that this p.d.f. is the same as the marginal bivariate p.d.f. of (X_1, X_2) found in Eq. (3.7.5).

From this marginal bivariate p.d.f., we can evaluate probabilities involving X_1 and X_2, such as $\Pr(X_1 + X_2 < 4)$. We have

$$\Pr(X_1 + X_2 < 4) = \int_0^4 \int_0^{4-x_2} \frac{4}{(2 + x_1 + x_2)^3}\, dx_1\, dx_2 = \frac{4}{9}.$$ ◀

Example 3.7.16 Service Times in a Queue. We can think of the random variable Z in Example 3.7.15 as the rate at which customers are served in the queue of Example 3.7.5. With this interpretation, it is useful to find the conditional distribution of the rate Z after we observe some of the service times such as X_1 and X_2.

For every value of z, the conditional p.d.f. of Z given $X_1 = x_1$ and $X_2 = x_2$ is

$$g_0(z|x_1, x_2) = \frac{f(z, x_1, x_2)}{f_{12}(x_1, x_2)} = \begin{cases} \frac{1}{2}(2 + x_1 + x_2)^3 z^2 e^{-z(2+x_1+x_2)} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.13}$$

Finally, we shall evaluate $\Pr(Z \le 1|X_1 = 1,\ X_2 = 4)$. We have

$$\Pr(Z \le 1|X_1 = 1,\ X_2 = 4) = \int_0^1 g_0(z|1, 4)\, dz = \int_0^1 171.5 z^2 e^{-7z}\, dz = 0.9704. \quad \blacktriangleleft$$

Law of Total Probability and Bayes' Theorem Example 3.7.15 contains an example of the multivariate version of the law of total probability, while Example 3.7.16 contains an example of the multivariate version of Bayes' theorem. The proofs of the general versions are straightforward consequences of Definition 3.7.7.

Theorem 3.7.5 Multivariate Law of Total Probability and Bayes' Theorem. Assume the conditions and notation given in Definition 3.7.7. If $\boldsymbol{Z}$ has a continuous joint distribution, the marginal p.d.f. of $\boldsymbol{Y}$ is

$$f_1(\boldsymbol{y}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-k} g_1(\boldsymbol{y}|\boldsymbol{z})\, f_2(\boldsymbol{z})\, d\boldsymbol{z}, \tag{3.7.14}$$

and the conditional p.d.f. of $\boldsymbol{Z}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is

$$g_2(\boldsymbol{z}|\boldsymbol{y}) = \frac{g_1(\boldsymbol{y}|\boldsymbol{z})\, f_2(\boldsymbol{z})}{f_1(\boldsymbol{y})}. \tag{3.7.15}$$

If $\boldsymbol{Z}$ has a discrete joint distribution, then the multiple integral in (3.7.14) must be replaced by a multiple summation. If $\boldsymbol{Z}$ has a mixed joint distribution, the multiple integral must be replaced by integration over those coordinates with continuous distributions and summation over those coordinates with discrete distributions. ■

Conditionally Independent Random Variables In Examples 3.7.15 and 3.7.16, $\boldsymbol{Z}$ is the single random variable Z and $\boldsymbol{Y} = (X_1, X_2)$. These examples also illustrate the use of conditionally independent random variables. That is, X_1 and X_2 are conditionally independent given $Z = z$ for all $z > 0$. In Example 3.7.16, we said that Z was the rate at which customers were served. When this rate is unknown, it is a major source of uncertainty. Partitioning the sample space by the values of the rate Z and then conditioning on each value of Z removes a major source of uncertainty for part of the calculation.

In general, conditional independence for random variables is similar to conditional independence for events.

Definition 3.7.8 Conditionally Independent Random Variables. Let $\boldsymbol{Z}$ be a random vector with joint p.f., p.d.f., or p.f./p.d.f. $f_0(\boldsymbol{z})$. Several random variables $X_1, \ldots, X_n$ are *conditionally independent given* $\boldsymbol{Z}$ if, for all $\boldsymbol{z}$ such that $f_0(\boldsymbol{z}) > 0$, we have

$$g(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{n} g_i(x_i|\boldsymbol{z}),$$

where $g(\boldsymbol{x}|\boldsymbol{z})$ stands for the conditional multivariate p.f., p.d.f., or p.f./p.d.f. of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$ and $g_i(x_i|\boldsymbol{z})$ stands for the conditional univariate p.f. or p.d.f. of X_i given $\boldsymbol{Z} = \boldsymbol{z}$.

In Example 3.7.15, $g_i(x_i|z) = ze^{-zx_i}$ for $x_i > 0$ and $i = 1, 2$.

Example 3.7.17 A Clinical Trial. In Example 3.7.8, the joint p.f./p.d.f. given there was constructed by assuming that $X_1, \ldots, X_m$ were conditionally independent given $P = p$ each with the same conditional p.f., $g_i(x_i|p) = p^{x_i}(1-p)^{1-x_i}$ for $x_i \in \{0, 1\}$ and that P had the uniform distribution on the interval $[0, 1]$. These assumptions produce, in the notation of Definition 3.7.8,

$$g(\boldsymbol{x}|p) = \begin{cases} p^{x_1+\cdots+x_m}(1-p)^{40-x_1-\cdots-x_m} & \text{for all } x_i \in \{0, 1\} \text{ and } 0 \le p \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

for $0 \le p \le 1$. Combining this with the marginal p.d.f. of P, $f_2(p) = 1$ for $0 \le p \le 1$ and 0 otherwise, we get the joint p.f./p.d.f. given in Example 3.7.8. ◀

Conditional Versions of Past and Future Theorems We mentioned earlier that conditional distributions behave just like distributions. Hence, all theorems that we have proven and will prove in the future have conditional versions. For example, the law of total probability in Eq. (3.7.14) has the following version conditional on another random vector $\boldsymbol{W} = \boldsymbol{w}$:

$$f_1(\boldsymbol{y}|\boldsymbol{w}) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-k} g_1(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w}) f_2(\boldsymbol{z}|\boldsymbol{w})\, d\boldsymbol{z}, \tag{3.7.16}$$

where $f_1(\boldsymbol{y}|\boldsymbol{w})$ stands for the conditional p.d.f., p.f., or p.f./p.d.f. of $\boldsymbol{Y}$ given $\boldsymbol{W} = \boldsymbol{w}$, $g_1(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w})$ stands for the conditional p.d.f., p.f., or p.f./p.d.f. of $\boldsymbol{Y}$ given $(\boldsymbol{Z}, \boldsymbol{W}) = (\boldsymbol{z}, \boldsymbol{w})$, and $f_2(\boldsymbol{z}|\boldsymbol{w})$ stands for the conditional p.d.f. of $\boldsymbol{Z}$ given $\boldsymbol{W} = \boldsymbol{w}$. Using the same notation, the conditional version of Bayes' theorem is

$$g_2(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}) = \frac{g_1(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{w}) f_2(\boldsymbol{z}|\boldsymbol{w})}{f_1(\boldsymbol{y}|\boldsymbol{w})}. \tag{3.7.17}$$

Example 3.7.18 Conditioning on Random Variables in Sequence. In Example 3.7.15, we found the conditional p.d.f. of Z given $(X_1, X_2) = (x_1, x_2)$. Suppose now that there are three more observations available, X_3, X_4, and X_5, and suppose that all of $X_1, \ldots, X_5$ are conditionally i.i.d. given $Z = z$ with p.d.f. $g(x|z)$. We shall use the conditional version of Bayes' theorem to compute the conditional p.d.f. of Z given $(X_1, \ldots, X_5) = (x_1, \ldots, x_5)$. First, we shall find the conditional p.d.f. $g_{345}(x_3, x_4, x_5|x_1, x_2, z)$ of $\boldsymbol{Y} = (X_3, X_4, X_5)$ given $Z = z$ and $\boldsymbol{W} = (X_1, X_2) = (x_1, x_2)$. We shall use the notation for p.d.f.'s in the discussion immediately preceding this example. Since $X_1, \ldots, X_5$ are conditionally i.i.d. given Z, we have that $g_1(\boldsymbol{y}|z, \boldsymbol{w})$ does not depend on $\boldsymbol{w}$. In fact,

$$g_1(\boldsymbol{y}|z, \boldsymbol{w}) = g(x_3|z)g(x_4|z)g(x_5|z) = z^3 e^{-z(x_3+x_4+x_5)},$$

for $x_3, x_4, x_5 > 0$. We also need the conditional p.d.f. of Z given $\boldsymbol{W} = \boldsymbol{w}$, which was calculated in Eq. (3.7.13), and we now denote it

$$f_2(z|\boldsymbol{w}) = \frac{1}{2}(2 + x_1 + x_2)^3 z^2 e^{-z(2+x_1+x_2)}.$$

Finally, we need the conditional p.d.f. of the last three observations given the first two. This was calculated in Example 3.7.14, and we now denote it

$$f_1(\boldsymbol{y}|\boldsymbol{w}) = \frac{60(2 + x_1 + x_2)^3}{(2 + x_1 + \cdots + x_5)^6}.$$

Now combine these using Bayes' theorem (3.7.17) to obtain

$$g_2(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}) = \frac{z^3 e^{-z(x_3+x_4+x_5)} \frac{1}{2}(2 + x_1 + x_2)^3 z^2 e^{-z(2+x_1+x_2)}}{\frac{60(2 + x_1 + x_2)^3}{(2 + x_1 + \cdots + x_5)^6}}$$

$$= \frac{1}{120}(2 + x_1 + \cdots + x_5)^6 z^5 e^{-z(2+x_1+\cdots+x_5)},$$

for $z > 0$. ◀

Note: Simple Rule for Creating Conditional Versions of Results. If you ever wish to determine the conditional version given $\boldsymbol{W} = \boldsymbol{w}$ of a result that you have proven, here is a simple method. Just add "conditional on $\boldsymbol{W} = \boldsymbol{w}$" to every probabilistic statement in the result. This includes all probabilities, c.d.f.'s, quantiles, names of distributions, p.d.f.'s, p.f.'s, and so on. It also includes all future probabilistic concepts that we introduce in later chapters (such as expected values and variances in Chapter 4).

Note: Independence is a Special Case of Conditional Independence. Let $X_1, \ldots, X_n$ be independent random variables, and let W be a constant random variable. That is, there is a constant c such that $\Pr(W = c) = 1$. Then $X_1, \ldots, X_n$ are also conditionally independent given $W = c$. The proof is straightforward and is left to the reader (Exercise 15). This result is not particularly interesting in its own right. Its value is the following: If we prove a result for conditionally independent random variables or conditionally i.i.d. random variables, then the same result will hold for independent random variables or i.i.d. random variables as the case may be.

Histograms

Example 3.7.19

Rate of Service. In Examples 3.7.5 and 3.7.6, we considered customers arriving at a queue and being served. Let Z stand for the rate at which customers were served, and we let $X_1, X_2, \ldots$ stand for the times that the successive customers requrired for service. Assume that $X_1, X_2, \ldots$ are conditionally i.i.d. given $Z = z$ with p.d.f.

$$g(x|z) = \begin{cases} ze^{-zx} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.7.18}$$

This is the same as (3.7.12) from Example 3.7.15. In that example, we modeled Z as a random variable with p.d.f. $f_0(z) = 2\exp(-2z)$ for $z > 0$. In this example, we shall assume that $X_1, \ldots, X_n$ will be observed for some large value n, and we want to think about what these observations tell us about Z. To be specific, suppose that we observe $n = 100$ service times. The first 10 times are listed here:

1.39, 0.61, 2.47, 3.35, 2.56, 3.60, 0.32, 1.43, 0.51, 0.94.

The smallest and largest observed service times from the entire sample are 0.004 and 9.60, respectively. It would be nice to have a graphical display of the entire sample of $n = 100$ service times without having to list them separately. ◀

The histogram, defined below, is a graphical display of a collection of numbers. It is particularly useful for displaying the observed values of a collection of random variables that have been modeled as conditionally i.i.d.

Definition 3.7.9

Histogram. Let $x_1, \ldots, x_n$ be a collection of numbers that all lie between two values $a < b$. That is, $a \le x_i \le b$ for all $i = 1, \ldots, n$. Choose some integer $k \ge 1$ and divide the interval $[a, b]$ into k equal-length subintervals of length $(b - a)/k$. For each subinterval, count how many of the numbers $x_1, \ldots, x_n$ are in the subinterval. Let c_i be the count for subinterval i for $i = 1, \ldots, k$. Choose a number $r > 0$. (Typically, $r = 1$ or $r = n$ or $r = n(b - a)/k$.) Draw a two-dimensional graph with the horizonal axis running from a to b. For each subinterval $i = 1, \ldots, k$ draw a rectangular bar of width $(b - a)/k$ and height equal to c_i/r over the midpoint of the ith interval. Such a graph is called a *histogram*.

The choice of the number r in the definition of histogram depends on what one wishes to be displayed on the vertical axis. The shape of the histogram is identical regardless of what value one chooses for r. With $r = 1$, the height of each bar is the raw count for each subinterval, and counts are displayed on the vertical axis. With $r = n$, the height of each bar is the proportion of the set of numbers in each subinterval, and the vertical axis displays proportions. With $r = n(b - a)/k$, the area of each bar is the proportion of the set of numbers in each subinterval.

Example 3.7.20

Rate of Service. The $n = 100$ observed service times in Example 3.7.19 all lie between 0 and 10. It is convenient, in this example, to draw a histogram with horizontal axis running from 0 to 10 and divided into 10 subintervals of length 1 each. Other choices are possible, but this one will do for illustration. Figure 3.22 contains the histogram of the 100 observed service times with $r = 100$. One sees that the numbers of observed service times in the subintervals decrease as the center of the subinterval increses. This matches the behavior of the conditional p.d.f. $g(x|z)$ of the service times as a function of x for fixed z. ◀

Histograms are useful as more than just graphical displays of large sets of numbers. After we see the law of large numbers (Theorem 6.2.4), we can show that the

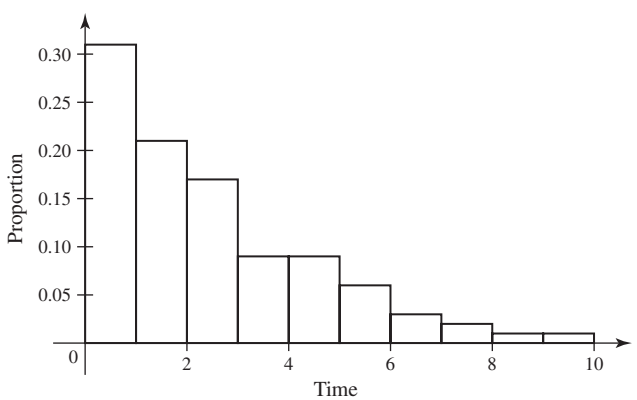

Figure 3.22 Histogram of service times for Example 3.7.20 with $a = 0$, $b = 10$, $k = 10$, and $r = 100$.

histogram of a large (conditionally) i.i.d. sample of continuous random variables is an approximation to the (conditional) p.d.f. of the random variables in the sample, so long as one uses the third choice of r, namely, $r = n(b-a)/k$.

Note: More General Histograms. Sometimes it is convenient to divide the range of the numbers to be plotted in a histogram into unequal-length subintervals. In such a case, one would typically let the height of each bar be c_i/r_i, where c_i is the raw count and r_i is proportional to the length of the ith subinterval. In this way, the area of each bar is still proportional to the count or proportion in each subinterval.

Summary

A finite collection of random variables is called a random vector. We have defined joint distributions for arbitrary random vectors. Every random vector has a joint c.d.f. Continuous random vectors have a joint p.d.f. Discrete random vectors have a joint p.f. Mixed distribution random vectors have a joint p.f./p.d.f. The coordinates of an n-dimensional random vector $\boldsymbol{X}$ are independent if the joint p.f., p.d.f., or p.f./p.d.f. $f(\boldsymbol{x})$ factors into $\prod_{i=1}^{n} f_i(x_i)$.

We can compute marginal distributions of subvectors of a random vector, and we can compute the conditional distribution of one subvector given the rest of the vector. We can construct a joint distribution for a random vector by piecing together a marginal distribution for part of the vector and a conditional distribution for the rest given the first part. There are versions of Bayes' theorem and the law of total probability for random vectors.

An n-dimensional random vector $\boldsymbol{X}$ has coordinates that are conditionally independent given $\boldsymbol{Z}$ if the conditional p.f., p.d.f., or p.f./p.d.f. $g(\boldsymbol{x}|\boldsymbol{z})$ of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$ factors into $\prod_{i=1}^{n} g_i(x_i|\boldsymbol{z})$. There are versions of Bayes' theorem, the law of total probability, and all future theorems about random variables and random vectors conditional on an arbitrary additional random vector.

Exercises

1. Suppose that three random variables X_1, X_2, and X_3 have a continuous joint distribution with the following joint p.d.f.: $f(x_1, x_2, x_3) =$

$$\begin{cases} c(x_1 + 2x_2 + 3x_3) & \text{for } 0 \le x_i \le 1 \ \ (i = 1, 2, 3), \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c;
(b) the marginal joint p.d.f. of X_1 and X_3; and
(c) $\Pr\left(X_3 < \frac{1}{2} \middle| X_1 = \frac{1}{4},\ X_2 = \frac{3}{4}\right)$.

2. Suppose that three random variables X_1, X_2, and X_3 have a mixed joint distribution with p.f./p.d.f.:

$$f(x_1, x_2, x_3) = \begin{cases} c x_1^{1+x_2+x_3}(1-x_1)^{3-x_2-x_3} & \text{if } 0 < x_1 < 1 \\ & \quad \text{and } x_2, x_3 \in \{0, 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

(Notice that X_1 has a continuous distribution and X_2 and X_3 have discrete distributions.) Determine **(a)** the value of the constant c; **(b)** the marginal joint p.f. of X_2 and X_3; and **(c)** the conditional p.d.f. of X_1 given $X_2 = 1$ and $X_3 = 1$.

3. Suppose that three random variables X_1, X_2, and X_3 have a continuous joint distribution with the following joint p.d.f.: $f(x_1, x_2, x_3) =$

$$\begin{cases} c e^{-(x_1+2x_2+3x_3)} & \text{for } x_i > 0 \ \ (i = 1, 2, 3), \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the value of the constant c; **(b)** the marginal joint p.d.f. of X_1 and X_3; and **(c)** $\Pr(X_1 < 1 | X_2 = 2,\ X_3 = 1)$.

4. Suppose that a point (X_1, X_2, X_3) is chosen at random, that is, in accordance with the uniform p.d.f., from the following set S:

$$S = \{(x_1, x_2, x_3): 0 \le x_i \le 1 \text{ for } i = 1, 2, 3\}.$$

Determine:

a. $\Pr\left[\left(X_1 - \frac{1}{2}\right)^2 + \left(X_2 - \frac{1}{2}\right)^2 + \left(X_3 - \frac{1}{2}\right)^2 \leq \frac{1}{4}\right]$

b. $\Pr(X_1^2 + X_2^2 + X_3^2 \leq 1)$

5. Suppose that an electronic system contains n components that function independently of each other and that the probability that component i will function properly is p_i $(i = 1, \ldots, n)$. It is said that the components are connected *in series* if a necessary and sufficient condition for the system to function properly is that all n components function properly. It is said that the components are connected *in parallel* if a necessary and sufficient condition for the system to function properly is that at least one of the n components functions properly. The probability that the system will function properly is called the *reliability* of the system. Determine the reliability of the system, **(a)** assuming that the components are connected in series, and **(b)** assuming that the components are connected in parallel.

6. Suppose that the n random variables $X_1 \ldots, X_n$ form a random sample from a discrete distribution for which the p.f. is f. Determine the value of $\Pr(X_1 = X_2 = \cdots = X_n)$.

7. Suppose that the n random variables $X_1, \ldots, X_n$ form a random sample from a continuous distribution for which the p.d.f. is f. Determine the probability that at least k of these n random variables will lie in a specified interval $a \leq x \leq b$.

8. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{1}{n!}x^n e^{-x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that for any given value $X = x$ $(x > 0)$, the n random variables $Y_1, \ldots, Y_n$ are i.i.d. and the conditional p.d.f. g of each of them is as follows:

$$g(y|x) = \begin{cases} \frac{1}{x} & \text{for } 0 < y < x, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the marginal joint p.d.f. of $Y_1, \ldots, Y_n$ and **(b)** the conditional p.d.f. of X for any given values of $Y_1, \ldots, Y_n$.

9. Let X be a random variable with a continuous distribution. Let $X_1 = X_2 = X$.

a. Prove that both X_1 and X_2 have a continuous distribution.

b. Prove that $\boldsymbol{X} = (X_1, X_2)$ does not have a continuous joint distribution.

10. Return to the situation described in Example 3.7.18. Let $\boldsymbol{X} = (X_1, \ldots, X_5)$ and compute the conditional p.d.f. of Z given $\boldsymbol{X} = \boldsymbol{x}$ directly in one step, as if all of $\boldsymbol{X}$ were observed at the same time.

11. Suppose that $X_1, \ldots, X_n$ are independent. Let $k < n$ and let $i_1, \ldots, i_k$ be distinct integers between 1 and n. Prove that $X_{i_1}, \ldots, X_{i_k}$ are independent.

12. Let $\boldsymbol{X}$ be a random vector that is split into three parts, $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W})$. Suppose that $\boldsymbol{X}$ has a continuous joint distribution with p.d.f. $f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w})$. Let $g_1(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{w})$ be the conditional p.d.f. of $(\boldsymbol{Y}, \boldsymbol{Z})$ given $\boldsymbol{W} = \boldsymbol{w}$, and let $g_2(\boldsymbol{y}|\boldsymbol{w})$ be the conditional p.d.f. of $\boldsymbol{Y}$ given $\boldsymbol{W} = \boldsymbol{w}$. Prove that $g_2(\boldsymbol{y}|\boldsymbol{w}) = \int g_1(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{w})\, d\boldsymbol{z}$.

13. Let X_1, X_2, X_3 be conditionally independent given $Z = z$ for all z with the conditional p.d.f. $g(x|z)$ in Eq. (3.7.12). Also, let the marginal p.d.f. of Z be f_0 in Eq. (3.7.11). Prove that the conditional p.d.f. of X_3 given $(X_1, X_2) = (x_1, x_2)$ is $\int_0^\infty g(x_3|z)g_0(z|x_1, x_2)\, dz$, where g_0 is defined in Eq. (3.7.13). (You can prove this even if you cannot compute the integral in closed form.)

14. Consider the situation described in Example 3.7.14. Suppose that $X_1 = 5$ and $X_2 = 7$ are observed.

a. Compute the conditional p.d.f. of X_3 given $(X_1, X_2) = (5, 7)$. (You may use the result stated in Exercise 12.)

b. Find the conditional probability that $X_3 > 3$ given $(X_1, X_2) = (5, 7)$ and compare it to the value of $\Pr(X_3 > 3)$ found in Example 3.7.9. Can you suggest a reason why the conditional probability should be higher than the marginal probability?

15. Let $X_1, \ldots, X_n$ be independent random variables, and let W be a random variable such that $\Pr(W = c) = 1$ for some constant c. Prove that $X_1, \ldots, X_n$ are conditionally independent given $W = c$.

3.8 Functions of a Random Variable

Often we find that after we compute the distribution of a random variable X, we really want the distribution of some function of X. For example, if X is the rate at which customers are served in a queue, then $1/X$ is the average waiting time. If we have the distribution of X, we should be able to determine the distribution of $1/X$ or of any other function of X. How to do that is the subject of this section.

Random Variable with a Discrete Distribution

Example 3.8.1 Distance from the Middle. Let X have the uniform distribution on the integers $1, 2, \ldots, 9$. Suppose that we are interested in how far X is from the middle of the distribution, namely, 5. We could define $Y = |X - 5|$ and compute probabilities such as $\Pr(Y = 1) = \Pr(X \in \{4, 6\}) = 2/9$. ◀

Example 3.8.1 illustrates the general procedure for finding the distribution of a function of a discrete random variable. The general result is straightforward.

Theorem 3.8.1 Function of a Discrete Random Variable. Let X have a discrete distribution with p.f. f, and let $Y = r(X)$ for some function of r defined on the set of possible values of X. For each possible value y of Y, the p.f. g of Y is

$$g(y) = \Pr(Y = y) = \Pr[r(X) = y] = \sum_{x:r(x)=y} f(x). \quad \blacksquare$$

Example 3.8.2 Distance from the Middle. The possible values of Y in Example 3.8.1 are 0, 1, 2, 3, and 4. We see that $Y = 0$ if and only if $X = 5$, so $g(0) = f(5) = 1/9$. For all other values of Y, there are two values of X that give that value of Y. For example, $\{Y = 4\} = \{X = 1\} \cup \{X = 9\}$. So, $g(y) = 2/9$ for $y = 1, 2, 3, 4$. ◀

Random Variable with a Continuous Distribution

If a random variable X has a continuous distribution, then the procedure for deriving the probability distribution of a function of X differs from that given for a discrete distribution. One way to proceed is by direct calculation as in Example 3.8.3.

Example 3.8.3 Average Waiting Time. Let Z be the rate at which customers are served in a queue, and suppose that Z has a continuous c.d.f. F. The average waiting time is $Y = 1/Z$. If we want to find the c.d.f. G of Y, we can write

$$G(y) = \Pr(Y \le y) = \Pr\left(\frac{1}{Z} \le y\right) = \Pr\left(Z \ge \frac{1}{y}\right) = \Pr\left(Z > \frac{1}{y}\right) = 1 - F\left(\frac{1}{y}\right),$$

where the fourth equality follows from the fact that Z has a continuous distribution so that $\Pr(Z = 1/y) = 0$. ◀

In general, suppose that the p.d.f. of X is f and that another random variable is defined as $Y = r(X)$. For each real number y, the c.d.f. $G(y)$ of Y can be derived as follows:

$$G(y) = \Pr(Y \le y) = \Pr[r(X) \le y]$$
$$= \int_{\{x:r(x)\le y\}} f(x)\,dx.$$

If the random variable Y also has a continuous distribution, its p.d.f. g can be obtained from the relation

$$g(y) = \frac{dG(y)}{dy}.$$

This relation is satisfied at every point y at which G is differentiable.

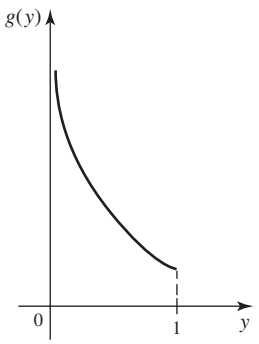

Figure 3.23 The p.d.f. of $Y = X^2$ in Example 3.8.4.

Example 3.8.4 Deriving the p.d.f. of X^2 when X Has a Uniform Distribution. Suppose that X has the uniform distribution on the interval $[-1, 1]$, so

$$f(x) = \begin{cases} 1/2 & \text{for } -1 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the p.d.f. of the random variable $Y = X^2$.

Since $Y = X^2$, then Y must belong to the interval $0 \le Y \le 1$. Thus, for each value of Y such that $0 \le y \le 1$, the c.d.f. $G(y)$ of Y is

$$\begin{aligned} G(y) &= \Pr(Y \le y) = \Pr(X^2 \le y) \\ &= \Pr(-y^{1/2} \le X \le y^{1/2}) \\ &= \int_{-y^{1/2}}^{y^{1/2}} f(x)\, dx = y^{1/2}. \end{aligned}$$

For $0 < y < 1$, it follows that the p.d.f. $g(y)$ of Y is

$$g(y) = \frac{dG(y)}{dy} = \frac{1}{2y^{1/2}}.$$

This p.d.f. of Y is sketched in Fig. 3.23. It should be noted that although Y is simply the square of a random variable with a uniform distribution, the p.d.f. of Y is unbounded in the neighborhood of $y = 0$. ◀

Linear functions are very useful transformations, and the p.d.f. of a linear function of a continuous random variable is easy to derive. The proof of the following result is left to the reader in Exercise 5.

Theorem 3.8.2 Linear Function. Suppose that X is a random variable for which the p.d.f. is f and that $Y = aX + b$ $(a \ne 0)$. Then the p.d.f. of Y is

$$g(y) = \frac{1}{|a|} f\left(\frac{y - b}{a}\right) \quad \text{for } -\infty < y < \infty, \tag{3.8.1}$$

and 0 otherwise. ■

The Probability Integral Transformation

Example 3.8.5 Let X be a continuous random variable with p.d.f. $f(x) = \exp(-x)$ for $x > 0$ and 0 otherwise. The c.d.f. of X is $F(x) = 1 - \exp(-x)$ for $x > 0$ and 0 otherwise. If we let

F be the function r in the earlier results of this section, we can find the distribution of $Y = F(X)$. The c.d.f. or Y is, for $0 < y < 1$,

$$G(y) = \Pr(Y \le y) = \Pr(1 - \exp(-X) \le y) = \Pr(X \le -\log(1-y))$$
$$= F(-\log(1-y)) = 1 - \exp(-[-\log(1-y)]) = y,$$

which is the c.d.f. of the uniform distribution on the interval [0, 1]. It follows that Y has the uniform distribution on the interval [0, 1]. ◀

The result in Example 3.8.5 is quite general.

Theorem 3.8.3 Probability Integral Transformation. Let X have a continuous c.d.f. F, and let $Y = F(X)$. (This transformation from X to Y is called the *probability integral transformation*.) The distribution of Y is the uniform distribution on the interval [0, 1].

Proof First, because F is the c.d.f. of a random variable, then $0 \le F(x) \le 1$ for $-\infty < x < \infty$. Therefore, $\Pr(Y < 0) = \Pr(Y > 1) = 0$. Since F is continuous, the set of x such that $F(x) = y$ is a nonempty closed and bounded interval $[x_0, x_1]$ for each y in the interval (0, 1). Let $F^{-1}(y)$ denote the lower endpoint x_0 of this interval, which was called the y quantile of F in Definition 3.3.2. In this way, $Y \le y$ if and only if $X \le x_1$. Let G denote the c.d.f. of Y. Then

$$G(y) = \Pr(Y \le y) = \Pr(X \le x_1) = F(x_1) = y.$$

Hence, $G(y) = y$ for $0 < y < 1$. Because this function is the c.d.f. of the uniform distribution on the interval [0, 1], this uniform distribution is the distribution of Y. ■

Because $\Pr(X = F^{-1}(Y)) = 1$ in the proof of Theorem 3.8.3, we have the following corollary.

Corollary 3.8.1 Let Y have the uniform distribution on the interval [0, 1], and let F be a continuous c.d.f. with quantile function F^{-1}. Then $X = F^{-1}(Y)$ has c.d.f. F. ■

Theorem 3.8.3 and its corollary give us a method for transforming an arbitrary continuous random variable X into another random variable Z with any desired continuous distribution. To be specific, let X have a continuous c.d.f. F, and let G be another continuous c.d.f. Then $Y = F(X)$ has the uniform distribution on the interval [0, 1] according to Theorem 3.8.3, and $Z = G^{-1}(Y)$ has the c.d.f. G according to Corollary 3.8.1. Combining these, we see that $Z = G^{-1}[F(X)]$ has c.d.f. G.

Simulation

Pseudo-Random Numbers Most computer packages that do statistical analyses also produce what are called *pseudo-random numbers*. These numbers appear to have some of the properties that a random sample would have, even though they are generated by deterministic algorithms. The most fundamental of these programs are the ones that generate pseudo-random numbers that appear to have the uniform distribution on the interval [0, 1]. We shall refer to such functions as *uniform pseudo-random number generators*. The important features that a uniform pseudo-random number generator must have are the following.

The numbers that it produces need to be spread somewhat uniformly over the interval [0, 1], and they need to appear to be observed values of independent random

variables. This last feature is very complicated to word precisely. An example of a sequence that does *not* appear to be observations of independent random variables would be one that was perfectly evenly spaced. Another example would be one with the following behavior: Suppose that we look at the sequence $X_1, X_2, \ldots$ one at a time, and every time we find an $X_i > 0.5$, we write down the next number X_{i+1}. If the subsequence of numbers that we write down is not spread approximately uniformly over the interval [0, 1], then the original sequence does not look like observations of independent random variables with the uniform distribution on the interval [0, 1]. The reason is that the conditional distribution of X_{i+1} given that $X_i > 0.5$ is supposed to be uniform over the interval [0, 1], according to independence.

Generating Pseudo-Random Numbers Having a Specified Distribution A uniform pseudo-random number generator can be used to generate values of a random variable Y having any specified continuous c.d.f. G. If a random variable X has the uniform distribution on the interval [0, 1] and if the quantile function G^{-1} is defined as before, then it follows from Corollary 3.8.1 that the c.d.f. of the random variable $Y = G^{-1}(X)$ will be G. Hence, if a value of X is produced by a uniform pseudo-random number generator, then the corresponding value of Y will have the desired property. If n independent values $X_1, \ldots, X_n$ are produced by the generator, then the corresponding values $Y_1, \ldots, Y_n$ will appear to form a random sample of size n from the distribution with the c.d.f. G.

Example 3.8.6

Generating Independent Values from a Specified p.d.f. Suppose that a uniform pseudo-random number generator is to be used to generate three independent values from the distribution for which the p.d.f. g is as follows:

$$g(y) = \begin{cases} \frac{1}{2}(2 - y) & \text{for } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

For $0 < y < 2$, the c.d.f. G of the given distribution is

$$G(y) = y - \frac{y^2}{4}.$$

Also, for $0 < x < 1$, the inverse function $y = G^{-1}(x)$ can be found by solving the equation $x = G(y)$ for y. The result is

$$y = G^{-1}(x) = 2[1 - (1 - x)^{1/2}]. \tag{3.8.2}$$

The next step is to generate three uniform pseudo-random numbers x_1, x_2, and x_3 using the generator. Suppose that the three generated values are

$$x_1 = 0.4125, \quad x_2 = 0.0894, \quad x_3 = 0.8302.$$

When these values of x_1, x_2, and x_3 are substituted successively into Eq. (3.8.2), the values of y that are obtained are $y_1 = 0.47$, $y_2 = 0.09$, and $y_3 = 1.18$. These are then treated as the observed values of three independent random variables with the distribution for which the p.d.f. is g. ◀

If G is a general c.d.f., there is a method similar to Corollary 3.8.1 that can be used to transform a uniform random variable into a random variable with c.d.f. G. See Exercise 12 in this section. There are other computer methods for generating values from certain specified distributions that are faster and more accurate than using the quantile function. These topics are discussed in the books by Kennedy and

Gentle (1980) and Rubinstein (1981). Chapter 12 of this text contains techniques and examples that show how simulation can be used to solve statistical problems.

General Function In general, if X has a continuous distribution and if $Y = r(X)$, then it is not necessarily true that Y will also have a continuous distribution. For example, suppose that $r(x) = c$, where c is a constant, for all values of x in some interval $a \le x \le b$, and that $\Pr(a \le X \le b) > 0$. Then $\Pr(Y = c) > 0$. Since the distribution of Y assigns positive probability to the value c, this distribution cannot be continuous. In order to derive the distribution of Y in a case like this, the c.d.f. of Y must be derived by applying methods like those described above. For certain functions r, however, the distribution of Y will be continuous; and it will then be possible to derive the p.d.f. of Y directly without first deriving its c.d.f. We shall develop this case in detail at the end of this section.

Direct Derivation of the p.d.f. When r is One-to-One and Differentiable

Example 3.8.7 Average Waiting Time. Consider Example 3.8.3 again. The p.d.f. g of Y can be computed from $G(y) = 1 - F(1/y)$ because F and $1/y$ both have derivatives at enough places. We apply the chain rule for differentiation to obtain

$$g(y) = \frac{dG(y)}{dy} = -\left.\frac{dF(x)}{dx}\right|_{x=1/y}\left(-\frac{1}{y^2}\right) = f\left(\frac{1}{y}\right)\frac{1}{y^2},$$

except at $y = 0$ and at those values of y such that $F(x)$ is not differentiable at $x = 1/y$. ◀

Differentiable One-To-One Functions The method used in Example 3.8.7 generalizes to very arbitrary differentiable one-to-one functions. Before stating the general result, we should recall some properties of differentiable one-to-one functions from calculus. Let r be a differentiable one-to-one function on the open interval (a, b). Then r is either strictly increasing or strictly decreasing. Because r is also continuous, it will map the interval (a, b) to another open interval (α, β), called the *image of* (a, b) *under* r. That is, for each $x \in (a, b)$, $r(x) \in (\alpha, \beta)$, and for each $y \in (\alpha, \beta)$ there is $x \in (a, b)$ such that $y = r(x)$ and this y is unique because r is one-to-one. So the inverse s of r will exist on the interval (α, β), meaning that for $x \in (a, b)$ and $y \in (\alpha, \beta)$ we have $r(x) = y$ if and only if $s(y) = x$. The derivative of s will exist (possibly infinite), and it is related to the derivative of r by

$$\frac{ds(y)}{dy} = \left(\left.\frac{dr(x)}{dx}\right|_{x=s(y)}\right)^{-1}.$$

Theorem 3.8.4 Let X be a random variable for which the p.d.f. is f and for which $\Pr(a < X < b) = 1$. (Here, a and/or b can be either finite or infinite.) Let $Y = r(X)$, and suppose that $r(x)$ is differentiable and one-to-one for $a < x < b$. Let (α, β) be the image of the interval (a, b) under the function r. Let $s(y)$ be the inverse function of $r(x)$ for $\alpha < y < \beta$. Then the p.d.f. g of Y is

$$g(y) = \begin{cases} f[s(y)]\left|\dfrac{ds(y)}{dy}\right| & \text{for } \alpha < y < \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8.3}$$

Proof If r is increasing, then s is increasing, and for each $y \in (\alpha, \beta)$,

$$G(y) = \Pr(Y \le y) = \Pr[r(X) \le y] = \Pr[X \le s(y)] = F[s(y)].$$

It follows that G is differentiable at all y where both s is differentiable and where $F(x)$ is differentiable at $x = s(y)$. Using the chain rule for differentiation, it follows that the p.d.f. $g(y)$ for $\alpha < y < \beta$ will be

$$g(y) = \frac{dG(y)}{dy} = \frac{dF[s(y)]}{dy} = f[s(y)]\frac{ds(y)}{dy}. \tag{3.8.4}$$

Because s is increasing, $ds(y)/dy$ is positive; hence, it equals $|ds(y)/dy|$ and Eq. (3.8.4) implies Eq. (3.8.3). Similarly, if r is decreasing, then s is decreasing, and for each $y \in (\alpha, \beta)$,

$$G(y) = \Pr[r(X) \le y] = \Pr[X \ge s(y)] = 1 - F[s(y)].$$

Using the chain rule again, we differentiate G to get the p.d.f. of Y

$$g(y) = \frac{dG(y)}{dy} = -f[s(y)]\frac{ds(y)}{dy}. \tag{3.8.5}$$

Since s is strictly decreasing, $ds(y)/dy$ is negative so that $-ds(y)/dy$ equals $|ds(y)/dy|$. It follows that Eq. (3.8.5) implies Eq. (3.8.3). ■

Example 3.8.8

Microbial Growth. A popular model for populations of microscopic organisms in large environments is exponential growth. At time 0, suppose that v organisms are introduced into a large tank of water, and let X be the rate of growth. After time t, we would predict a population size of ve^{Xt}. Assume that X is unknown but has a continuous distribution with p.d.f.

$$f(x) = \begin{cases} 3(1-x)^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in the distribution of $Y = ve^{Xt}$ for known values of v and t. For concreteness, let $v = 10$ and $t = 5$, so that $r(x) = 10e^{5x}$.

In this example, $\Pr(0 < X < 1) = 1$ and r is a continuous and strictly increasing function of x for $0 < x < 1$. As x varies over the interval $(0, 1)$, it is found that $y = r(x)$ varies over the interval $(10, 10e^5)$. Furthermore, for $10 < y < 10e^5$, the inverse function is $s(y) = \log(y/10)/5$. Hence, for $10 < y < 10e^5$,

$$\frac{ds(y)}{dy} = \frac{1}{5y}.$$

It follows from Eq. (3.8.3) that $g(y)$ will be

$$g(y) = \begin{cases} \dfrac{3(1 - \log(y/10)/5)^2}{5y} & \text{for } 10 < y < 10e^5, \\ 0 & \text{otherwise.} \end{cases}$$

◄

Summary

We learned several methods for determining the distribution of a function of a random variable. For a random variable X with a continuous distribution having p.d.f. f, if r is strictly increasing or strictly decreasing with differentiable inverse s (i.e., $s(r(x)) = x$ and s is differentiable), then the p.d.f. of $Y = r(X)$ is $g(y) =$

$f(s(y))|ds(y)/dy|$. A special transformation allows us to transform a random variable X with the uniform distribution on the interval $[0, 1]$ into a random variable Y with an arbitrary continuous c.d.f. G by $Y = G^{-1}(X)$. This method can be used in conjunction with a uniform pseudo-random number generator to generate random variables with arbitrary continuous distributions.

Exercises

1. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, suppose that $Y = 1 - X^2$. Determine the p.d.f. of Y.

2. Suppose that a random variable X can have each of the seven values $-3, -2, -1, 0, 1, 2, 3$ with equal probability. Determine the p.f. of $Y = X^2 - X$.

3. Suppose that the p.d.f. of a random variable X is as follows:

$$f(x) = \begin{cases} \frac{1}{2}x & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Also, suppose that $Y = X(2 - X)$. Determine the c.d.f. and the p.d.f. of Y.

4. Suppose that the p.d.f. of X is as given in Exercise 3. Determine the p.d.f. of $Y = 4 - X^3$.

5. Prove Theorem 3.8.2. (*Hint:* Either apply Theorem 3.8.4 or first compute the c.d.f. seperately for $a > 0$ and $a < 0$.)

6. Suppose that the p.d.f. of X is as given in Exercise 3. Determine the p.d.f. of $Y = 3X + 2$.

7. Suppose that a random variable X has the uniform distribution on the interval $[0, 1]$. Determine the p.d.f. of **(a)** X^2, **(b)** $-X^3$, and **(c)** $X^{1/2}$.

8. Suppose that the p.d.f. of X is as follows:

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

Determine the p.d.f. of $Y = X^{1/2}$.

9. Suppose that X has the uniform distribution on the interval $[0, 1]$. Construct a random variable $Y = r(X)$ for which the p.d.f. will be

$$g(y) = \begin{cases} \frac{3}{8}y^2 & \text{for } 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

10. Let X be a random variable for which the p.d.f f is as given in Exercise 3. Construct a random variable $Y = r(X)$ for which the p.d.f. g is as given in Exercise 9.

11. Explain how to use a uniform pseudo-random number generator to generate four independent values from a distribution for which the p.d.f. is

$$g(y) = \begin{cases} \frac{1}{2}(2y + 1) & \text{for } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

12. Let F be an arbitrary c.d.f. (not necessarily discrete, not necessarily continuous, not necessarily either). Let F^{-1} be the quantile function from Definition 3.3.2. Let X have the uniform distribution on the interval $[0, 1]$. Define $Y = F^{-1}(X)$. Prove that the c.d.f. of Y is F. *Hint:* Compute $\Pr(Y \le y)$ in two cases. First, do the case in which y is the unique value of x such that $F(x) = F(y)$. Second, do the case in which there is an entire interval of x values such that $F(x) = F(y)$.

13. Let Z be the rate at which customers are served in a queue. Assume that Z has the p.d.f.

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

Find the p.d.f. of the average waiting time $T = 1/Z$.

14. Let X have the uniform distribution on the interval $[a, b]$, and let $c > 0$. Prove that $cX + d$ has the uniform distribution on the interval $[ca + d, cb + d]$.

15. Most of the calculation in Example 3.8.4 is quite general. Suppose that X has a continuous distribution with p.d.f. f. Let $Y = X^2$, and show that the p.d.f. of Y is

$$g(y) = \frac{1}{2y^{1/2}}[f(y^{1/2}) + f(-y^{1/2})].$$

16. In Example 3.8.4, the p.d.f. of $Y = X^2$ is much larger for values of y near 0 than for values of y near 1 despite the fact that the p.d.f. of X is flat. Give an intuitive reason why this occurs in this example.

17. An insurance agent sells a policy which has a \$100 deductible and a \$5000 cap. This means that when the policy holder files a claim, the policy holder must pay the first

\$100. After the first \$100, the insurance company pays the rest of the claim up to a maximum payment of \$5000. Any excess must be paid by the policy holder. Suppose that the dollar amount X of a claim has a continuous distribution with p.d.f. $f(x) = 1/(1+x)^2$ for $x > 0$ and 0 otherwise. Let Y be the amount that the insurance company has to pay on the claim.

a. Write Y as a function of X, i.e., $Y = r(X)$.

b. Find the c.d.f. of Y.

c. Explain why Y has neither a continuous nor a discrete distribution.

3.9 Functions of Two or More Random Variables

When we observe data consisting of the values of several random variables, we need to summarize the observed values in order to be able to focus on the information in the data. Summarizing consists of constructing one or a few functions of the random variables that capture the bulk of the information. In this section, we describe the techniques needed to determine the distribution of a function of two or more random variables.

Random Variables with a Discrete Joint Distribution

Example 3.9.1

Bull Market. Three different investment firms are trying to advertise their mutual funds by showing how many perform better than a recognized standard. Each company has 10 funds, so there are 30 in total. Suppose that the first 10 funds belong to the first firm, the next 10 to the second firm, and the last 10 to the third firm. Let $X_i = 1$ if fund i performs better than the standard and $X_i = 0$ otherwise, for $i = 1, \ldots, 30$. Then, we are interested in the three functions

$$Y_1 = X_1 + \cdots + X_{10},$$
$$Y_2 = X_{11} + \cdots + X_{20},$$
$$Y_3 = X_{21} + \cdots + X_{30}.$$

We would like to be able to determine the joint distribution of Y_1, Y_2, and Y_3 from the joint distribution of $X_1, \ldots, X_{30}$. ◀

The general method for solving problems like those of Example 3.9.1 is a straightforward extension of Theorem 3.8.1.

Theorem 3.9.1

Functions of Discrete Random Variables. Suppose that n random variables $X_1, \ldots, X_n$ have a discrete joint distribution for which the joint p.f. is f, and that m functions $Y_1, \ldots, Y_m$ of these n random variables are defined as follows:

$$Y_1 = r_1(X_1, \ldots, X_n),$$
$$Y_2 = r_2(X_1, \ldots, X_n),$$
$$\vdots$$
$$Y_m = r_m(X_1, \ldots, X_n).$$

For given values $y_1, \ldots, y_m$ of the m random variables $Y_1, \ldots, Y_m$, let A denote the set of all points $(x_1, \ldots, x_n)$ such that

$$\begin{aligned} r_1(x_1, \ldots, x_n) &= y_1, \\ r_2(x_1, \ldots, x_n) &= y_2, \\ &\vdots \\ r_m(x_1, \ldots, x_n) &= y_m. \end{aligned}$$

Then the value of the joint p.f. g of $Y_1, \ldots, Y_m$ is specified at the point $(y_1, \ldots, y_m)$ by the relation

$$g(y_1, \ldots, y_m) = \sum_{(x_1, \ldots, x_n) \in A} f(x_1, \ldots, x_n). \qquad \blacksquare$$

Example 3.9.2 Bull Market. Recall the situation in Example 3.9.1. Suppose that we want the joint p.f. g of (Y_1, Y_2, Y_3) at the point (3, 5, 8). That is, we want $g(3, 5, 8) = \Pr(Y_1 = 3, Y_2 = 5, Y_3 = 8)$. The set A as defined in Theorem 3.9.1 is

$$A = \{(x_1, \ldots, x_{30}) : x_1 + \cdots + x_{10} = 3, x_{11} + \cdots + x_{20} = 5, x_{21} + \cdots + x_{30} = 8\}.$$

Two of the points in the set A are

$$(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0),$$
$$(1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1).$$

A counting argument like those developed in Sec. 1.8 can be used to discover that there are

$$\binom{10}{3}\binom{10}{5}\binom{10}{8} = 1{,}360{,}800$$

points in A. Unless the joint distribution of $X_1, \ldots, X_{30}$ has some simple structure, it will be extremely tedious to compute $g(3, 5, 8)$ as well as most other values of g. For example, if all of the 2^{30} possible values of the vector $(X_1, \ldots, X_{30})$ are equally likely, then

$$g(3, 5, 8) = \frac{1{,}360{,}800}{2^{30}} = 1.27 \times 10^{-3}. \qquad \blacktriangleleft$$

The next result gives an important example of a function of discrete random variables.

Theorem 3.9.2 Binomial and Bernoulli Distributions. Assume that $X_1, \ldots, X_n$ are i.i.d. random variables having the Bernoulli distribution with parameter p. Let $Y = X_1 + \cdots + X_n$. Then Y has the binomial distribution with parameters n and p.

Proof It is clear that $Y = y$ if and only if exactly y of $X_1, \ldots, X_n$ equal 1 and the other $n - y$ equal 0. There are $\binom{n}{y}$ distinct possible values for the vector $(X_1, \ldots, X_n)$ that have y ones and $n - y$ zeros. Each such vector has probability $p^y(1 - p)^{n-y}$ of being observed; hence the probability that $Y = y$ is the sum of the probabilities of those vectors, namely, $\binom{n}{y}p^y(1 - p)^{n-y}$ for $y = 0, \ldots, n$. From Definition 3.1.7, we see that Y has the binomial distribution with parameters n and p. $\blacksquare$

Example 3.9.3 Sampling Parts. Suppose that two machines are producing parts. For $i = 1, 2$, the probability is p_i that machine i will produce a defective part, and we shall assume that all parts from both machines are independent. Assume that the first n_1 parts are produced by machine 1 and that the last n_2 parts are produced by machine 2,

with $n = n_1 + n_2$ being the total number of parts sampled. Let $X_i = 1$ if the ith part is defective and $X_i = 0$ otherwise for $i = 1, \ldots, n$. Define $Y_1 = X_1 + \cdots + X_{n_1}$ and $Y_2 = X_{n_1+1} + \cdots + X_n$. These are the total numbers of defective parts produced by each machine. The assumptions stated in the problem allow us to conclude that Y_1 and Y_2 are independent according to the note about separate functions of independent random variables on page 140. Furthermore, Theorem 3.9.2 says that Y_j has the binomial distribution with parameters n_j and p_j for $j = 1, 2$. These two marginal distributions, together with the fact that Y_1 and Y_2 are independent, give the entire joint distribution. So, for example, if g is the joint p.f. of Y_1 and Y_2, we can compute

$$g(y_1, y_2) = \binom{n_1}{y_1} p_1^{y_1}(1 - p_1)^{n_1 - y_1} \binom{n_2}{y_2} p_2^{y_2}(1 - p)^{n_2 - y_2},$$

for $y_1 = 0, \ldots, n_1$ and $y_2 = 0, \ldots, n_2$, while $g(y_1, y_2) = 0$ otherwise. There is no need to find a set A as in Example 3.9.2, because of the simplifying structure of the joint distribution of $X_1, \ldots, X_n$. ◀

Random Variables with a Continuous Joint Distribution

Example 3.9.4

Total Service Time. Suppose that the first two customers in a queue plan to leave together. Let X_i be the time it takes to serve customer i for $i = 1, 2$. Suppose also that X_1 and X_2 are independent random variables with common distribution having p.d.f. $f(x) = 2e^{-2x}$ for $x > 0$ and 0 otherwise. Since the customers will leave together, they are interested in the total time it takes to serve both of them, namely, $Y = X_1 + X_2$. We can now find the p.d.f. of Y.

For each y, let

$$A_y = \{(x_1, x_2) : x_1 + x_2 \le y\}.$$

Then $Y \le y$ if and only if $(X_1, X_2) \in A_y$. The set A_y is pictured in Fig. 3.24. If we let $G(y)$ denote the c.d.f. of Y, then, for $y > 0$,

$$\begin{aligned} G(y) = \Pr((X_1, X_2) \in A_y) &= \int_0^y \int_0^{y - x_2} 4e^{-2x_1 - 2x_2} dx_1 dx_2 \\ &= \int_0^y 2e^{-2x_2}\left[1 - e^{-2(y - x_2)}\right] dx_2 = \int_0^y \left[2e^{-2x_2} - 2e^{-2y}\right] dx_2 \\ &= 1 - e^{-2y} - 2ye^{-2y}. \end{aligned}$$

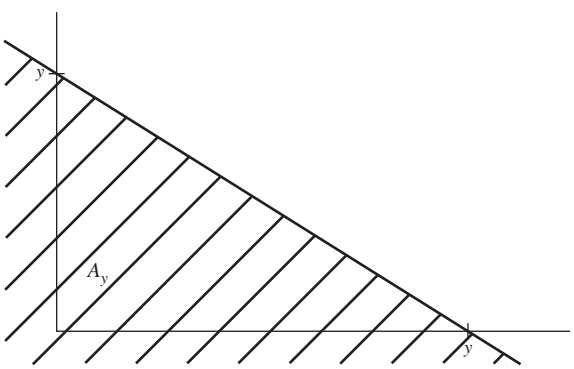

Figure 3.24 The set A_y in Example 3.9.4 and in the proof of Theorem 3.9.4.

Taking the derivative of $G(y)$ with respect to y, we get the p.d.f.

$$g(y) = \frac{d}{dy}\left[1 - e^{-2y} - ye^{-2y}\right] = 4ye^{-2y},$$

for $y > 0$ and 0 otherwise. ◀

The transformation in Example 3.9.4 is an example of a brute-force method that is always available for finding the distribution of a function of several random variables, however, it might be difficult to apply in individual cases.

Theorem 3.9.3 Brute-Force Distribution of a Function. Suppose that the joint p.d.f. of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is $f(\boldsymbol{x})$ and that $Y = r(\boldsymbol{X})$. For each real number y, define $A_y = \{\boldsymbol{x} : r(\boldsymbol{x}) \le y\}$. Then the c.d.f. $G(y)$ of Y is

$$G(y) = \int \cdots \int_{A_y} f(\boldsymbol{x})\, d\boldsymbol{x}. \tag{3.9.1}$$

Proof From the definition of c.d.f.,

$$G(y) = \Pr(Y \le y) = \Pr[r(\boldsymbol{X}) \le y] = \Pr(\boldsymbol{X} \in A_y),$$

which equals the right side of Eq. (3.9.1) by Definition 3.7.3. ■

If the distribution of Y also is continuous, then the p.d.f. of Y can be found by differentiating the c.d.f. $G(y)$.

A popular special case of Theorem 3.9.3 is the following.

Theorem 3.9.4 Linear Function of Two Random Variables. Let X_1 and X_2 have joint p.d.f. $f(x_1, x_2)$, and let $Y = a_1X_1 + a_2X_2 + b$ with $a_1 \ne 0$. Then Y has a continuous distribution whose p.d.f. is

$$g(y) = \int_{-\infty}^{\infty} f\left(\frac{y - b - a_2x_2}{a_1}, x_2\right)\frac{1}{|a_1|}dx_2. \tag{3.9.2}$$

Proof First, we shall find the c.d.f. G of Y whose derivative we will see is the function g in Eq. (3.9.2). For each y, let $A_y = \{(x_1, x_2) : a_1x_1 + a_2x_2 + b \le y\}$. The set A_y has the same general form as the set in Fig. 3.24. We shall write the integral over the set A_y with x_2 in the outer integral and x_1 in the inner integral. Assume that $a_1 > 0$. The other case is similar. According to Theorem 3.9.3,

$$G(y) = \int_{A_y}\int f(x_1, x_2)dx_1dx_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{(y-b-a_2x_2)/a_1} f(x_1, x_2)dx_1dx_2. \tag{3.9.3}$$

For the inner integral, perform the change of variable $z = a_1x_1 + a_2x_2 + b$ whose inverse is $x_1 = (z - b - a_2x_2)/a_1$, so that $dx_1 = dz/a_1$. The inner integral, after this change of variable, becomes

$$\int_{-\infty}^{y} f\left(\frac{z - b - a_2x_2}{a_1}, x_2\right)\frac{1}{a_1}dz.$$

We can now substitute this expression for the inner integral into Eq. (3.9.3):

$$\begin{aligned} G(y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{y} f\left(\frac{z - b - a_2x_2}{a_1}, x_2\right)\frac{1}{a_1}dzdx_2 \\ &= \int_{-\infty}^{y}\int_{-\infty}^{\infty} f\left(\frac{z - b - a_2x_2}{a_1}, x_2\right)\frac{1}{a_1}dx_2dz. \end{aligned} \tag{3.9.4}$$

Let $g(z)$ denote the inner integral on the far right side of Eq. (3.9.4). Then we have $G(y) = \int_{-\infty}^{y} g(z)dz$, whose derivative is $g(y)$, the function in Eq. (3.9.2). ■

The special case of Theorem 3.9.4 in which X_1 and X_2 are independent, $a_1 = a_2 = 1$, and $b = 0$ is called *convolution*.

Definition 3.9.1 Convolution. Let X_1 and X_2 be independent continuous random variables and let $Y = X_1 + X_2$. The distribution of Y is called the *convolution* of the distributions of X_1 and X_2. The p.d.f. of Y is sometimes called the convolution of the p.d.f.'s of X_1 and X_2.

If we let the p.d.f. of X_i be f_i for $i = 1, 2$ in Definition 3.9.1, then Theorem 3.9.4 (with $a_1 = a_2 = 1$ and $b = 0$) says that the p.d.f. of $Y = X_1 + X_2$ is

$$g(y) = \int_{-\infty}^{\infty} f_1(y - z) f_2(z) dz. \tag{3.9.5}$$

Equivalently, by switching the names of X_1 and X_2, we obtain the alternative form for the convolution:

$$g(y) = \int_{-\infty}^{\infty} f_1(z) f_2(y - z)\, dz. \tag{3.9.6}$$

The p.d.f. found in Example 3.9.4 is the special case of (3.9.5) with $f_1(x) = f_2(x) = 2e^{-2x}$ for $x > 0$ and 0 otherwise.

Example 3.9.5 An Investment Portfolio. Suppose that an investor wants to purchase both stocks and bonds. Let X_1 be the value of the stocks at the end of one year, and let X_2 be the value of the bonds at the end of one year. Suppose that X_1 and X_2 are independent. Let X_1 have the uniform distribution on the interval [1000, 4000], and let X_2 have the uniform distribution on the interval [800, 1200]. The sum $Y = X_1 + X_2$ is the value at the end of the year of the portfolio consisting of both the stocks and the bonds. We shall find the p.d.f. of Y. The function $f_1(z) f_2(y - z)$ in Eq. (3.9.6) is

$$f_1(z) f_2(y - z) = \begin{cases} 8.333 \times 10^{-7} & \text{for } 1000 \le z \le 4000 \\ & \quad \text{and } 800 \le y - z \le 1200, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9.7}$$

We need to integrate the function in Eq. (3.9.7) over z for each value of y to get the marginal p.d.f. of Y. It is helpful to look at a graph of the set of (y, z) pairs for which the function in Eq. (3.9.7) is positive. Figure 3.25 shows the region shaded. For $1800 < y \le 2200$, we must integrate z from 1000 to $y - 800$. For $2200 < y \le 4800$, we must integrate z from $y - 1200$ to $y - 800$. For $4800 < y < 5200$, we must integrate z from $y - 1200$ to 4000. Since the function in Eq. (3.9.7) is constant when it is positive, the integral equals the constant times the length of the interval of z values. So, the p.d.f. of Y is

$$g(y) = \begin{cases} 8.333 \times 10^{-7}(y - 1800) & \text{for } 1800 < y \le 2200, \\ 3.333 \times 10^{-4} & \text{for } 2200 < y \le 4800, \\ 8.333 \times 10^{-7}(5200 - y) & \text{for } 4800 < y < 5200, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

As another example of the brute-force method, we consider the largest and smallest observations in a random sample. These functions give an idea of how spread out the sample is. For example, meteorologists often report record high and low

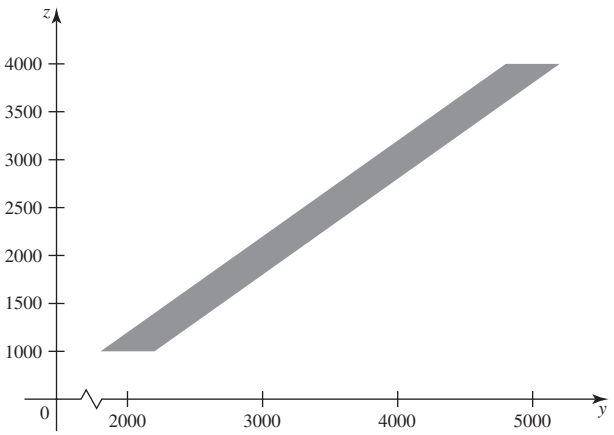

Figure 3.25 The region where the function in Eq. (3.9.7) is positive.

temperatures for specific days as well as record high and low rainfalls for months and years.

Example 3.9.6

Maximum and Minimum of a Random Sample. Suppose that $X_1, \ldots, X_n$ form a random sample of size n from a distribution for which the p.d.f. is f and the c.d.f. is F. The largest value Y_n and the smallest value Y_1 in the random sample are defined as follows:

$$Y_n = \max\{X_1, \ldots, X_n\},$$
$$Y_1 = \min\{X_1, \ldots, X_n\}. \tag{3.9.8}$$

Consider Y_n first. Let G_n stand for its c.d.f., and let g_n be its p.d.f. For every given value of y $(-\infty < y < \infty)$,

$$\begin{aligned} G_n(y) &= \Pr(Y_n \le y) = \Pr(X_1 \le y, X_2 \le y, \ldots, X_n \le y) \\ &= \Pr(X_1 \le y)\Pr(X_2 \le y) \cdots \Pr(X_n \le y) \\ &= F(y)F(y) \cdots F(y) = [F(y)]^n, \end{aligned}$$

where the third equality follows from the fact that the X_i are independent and the fourth follows from the fact that all of the X_i have the same c.d.f. F. Thus, $G_n(y) = [F(y)]^n$.

Now, g_n can be determined by differentiating the c.d.f. G_n. The result is

$$g_n(y) = n[F(y)]^{n-1} f(y) \quad \text{for } -\infty < y < \infty.$$

Next, consider Y_1 with c.d.f. G_1 and p.d.f. g_1. For every given value of y $(-\infty < y < \infty)$,

$$\begin{aligned} G_1(y) &= \Pr(Y_1 \le y) = 1 - \Pr(Y_1 > y) \\ &= 1 - \Pr(X_1 > y, X_2 > y, \ldots, X_n > y) \\ &= 1 - \Pr(X_1 > y)\Pr(X_2 > y) \cdots \Pr(X_n > y) \\ &= 1 - [1 - F(y)][1 - F(y)] \cdots [1 - F(y)] \\ &= 1 - [1 - F(y)]^n. \end{aligned}$$

Thus, $G_1(y) = 1 - [1 - F(y)]^n$.

Then g_1 can be determined by differentiating the c.d.f. G_1. The result is

$$g_1(y) = n[1 - F(y)]^{n-1} f(y) \quad \text{for } -\infty < y < \infty.$$

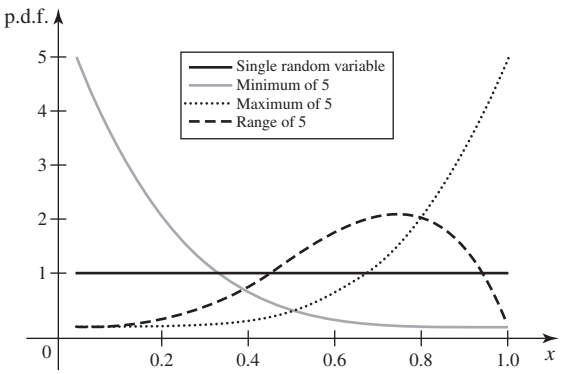

Figure 3.26 The p.d.f. of the uniform distribution on the interval [0, 1] together with the p.d.f.'s of the minimum and maximum of samples of size $n = 5$. The p.d.f. of the range of a sample of size $n = 5$ (see Example 3.9.7) is also included.

Figure 3.26 shows the p.d.f. of the uniform distribution on the interval [0, 1] together with the p.d.f.'s of Y_1 and Y_n for the case $n = 5$. It also shows the p.d.f. of $Y_5 - Y_1$, which will be derived in Example 3.9.7. Notice that the p.d.f. of Y_1 is highest near 0 and lowest near 1, while the opposite is true of the p.d.f. of Y_n, as one would expect.

Finally, we shall determine the joint distribution of Y_1 and Y_n. For every pair of values (y_1, y_n) such that $-\infty < y_1 < y_n < \infty$, the event $\{Y_1 \le y_1\} \cap \{Y_n \le y_n\}$ is the same as $\{Y_n \le y_n\} \cap \{Y_1 > y_1\}^c$. If G denotes the bivariate joint c.d.f. of Y_1 and Y_n, then

$$\begin{aligned} G(y_1, y_n) &= \Pr(Y_1 \le y_1 \text{ and } Y_n \le y_n) \\ &= \Pr(Y_n \le y_n) - \Pr(Y_n \le y_n \text{ and } Y_1 > y_1) \\ &= \Pr(Y_n \le y_n) \\ &\quad - \Pr(y_1 < X_1 \le y_n, y_1 < X_2 \le y_n, \ldots, y_1 < X_n \le y_n) \\ &= G_n(y_n) - \prod_{i=1}^{n} \Pr(y_1 < X_i \le y_n) \\ &= [F(y_n)]^n - [F(y_n) - F(y_1)]^n. \end{aligned}$$

The bivariate joint p.d.f. g of Y_1 and Y_n can be found from the relation

$$g(y_1, y_n) = \frac{\partial^2 G(y_1, y_n)}{\partial y_1 \partial y_n}.$$

Thus, for $-\infty < y_1 < y_n < \infty$,

$$g(y_1, y_n) = n(n-1)[F(y_n) - F(y_1)]^{n-2} f(y_1) f(y_n). \tag{3.9.9}$$

Also, for all other values of y_1 and y_n, $g(y_1, y_n) = 0$. ◀

A popular way to describe how spread out is a random sample is to use the distance from the minimum to the maximum, which is called the *range* of the random sample. We can combine the result from the end of Example 3.9.6 with Theorem 3.9.4 to find the p.d.f. of the range.

Example 3.9.7

The Distribution of the Range of a Random Sample. Consider the same situation as in Example 3.9.6. The random variable $W = Y_n - Y_1$ is called the *range* of the sample. The joint p.d.f. $g(y_1, y_n)$ of Y_1 and Y_n was presented in Eq. (3.9.9). We can now apply Theorem 3.9.4 with $a_1 = -1$, $a_2 = 1$, and $b = 0$ to get the p.d.f. h of W:

$$h(w) = \int_{-\infty}^{\infty} g(y_n - w, y_n) dy_n = \int_{-\infty}^{\infty} g(z, z + w) dz, \tag{3.9.10}$$

where, for the last equality, we have made the change of variable $z = y_n - w$. ◀

Here is a special case in which the integral of Eq. 3.9.10 can be computed in closed form.

Example 3.9.8 The Range of a Random Sample from a Uniform Distribution. Suppose that the n random variables $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, 1]$. We shall determine the p.d.f. of the range of the sample.

In this example,

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

Also, $F(x) = x$ for $0 < x < 1$. We can write $g(y_1, y_n)$ from Eq. (3.9.9) in this case as

$$g(y_1, y_n) = \begin{cases} n(n-1)(y_n - y_1)^{n-2} & \text{for } 0 < y_1 < y_n < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, in Eq. (3.9.10), $g(z, z + w) = 0$ unless $0 < w < 1$ and $0 < z < 1 - w$. For values of w and z satisfying these conditions, $g(z, w + z) = n(n-1)w^{n-2}$. The p.d.f. in Eq. (3.9.10) is then, for $0 < w < 1$,

$$h(w) = \int_0^{1-w} n(n-1)w^{n-2}\, dz = n(n-1)w^{n-2}(1 - w).$$

Otherwise, $h(w) = 0$. This p.d.f. is shown in Fig. 3.26 for the case $n = 5$. ◀

Direct Transformation of a Multivariate p.d.f.

Next, we state without proof a generalization of Theorem 3.8.4 to the case of several random variables. The proof of Theorem 3.9.5 is based on the theory of differentiable one-to-one transformations in advanced calculus.

Theorem 3.9.5 Multivariate Transformation. Let $X_1, \ldots, X_n$ have a continuous joint distribution for which the joint p.d.f. is f. Assume that there is a subset S of R^n such that $\Pr[(X_1, \ldots, X_n) \in S] = 1$. Define n new random variables $Y_1, \ldots, Y_n$ as follows:

$$\begin{aligned} Y_1 &= r_1(X_1, \ldots, X_n), \\ Y_2 &= r_2(X_1, \ldots, X_n), \\ &\vdots \\ Y_n &= r_n(X_1, \ldots, X_n), \end{aligned} \tag{3.9.11}$$

where we assume that the n functions $r_1, \ldots, r_n$ define a one-to-one differentiable transformation of S onto a subset T of R^n. Let the inverse of this transformation be given as follows:

$$\begin{aligned} x_1 &= s_1(y_1, \ldots, y_n), \\ x_2 &= s_2(y_1, \ldots, y_n), \\ &\vdots \\ x_n &= s_n(y_1, \ldots, y_n). \end{aligned} \tag{3.9.12}$$

Then the joint p.d.f. g of $Y_1, \ldots, Y_n$ is

$$g(y_1, \ldots, y_n) = \begin{cases} f(s_1, \ldots, s_n)|J| & \text{for } (y_1, \ldots, y_n) \in T, \\ 0 & \text{otherwise,} \end{cases} \tag{3.9.13}$$

where J is the determinant

$$J = \det \begin{bmatrix} \frac{\partial s_1}{\partial y_1} & \cdots & \frac{\partial s_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_n}{\partial y_1} & \cdots & \frac{\partial s_n}{\partial y_n} \end{bmatrix}$$

and $|J|$ denotes the absolute value of the determinant J. ■

Thus, the joint p.d.f. $g(y_1, \ldots, y_n)$ is obtained by starting with the joint p.d.f. $f(x_1, \ldots, x_n)$, replacing each value x_i by its expression $s_i(y_1, \ldots, y_n)$ in terms of $y_1, \ldots, y_n$, and then multiplying the result by $|J|$. This determinant J is called the *Jacobian* of the transformation specified by the equations in (3.9.12).

Note: The Jacobian Is a Generalization of the Derivative of the Inverse. Eqs. (3.8.3) and (3.9.13) are very similar. The former gives the p.d.f. of a single function of a single random variable. Indeed, if $n = 1$ in (3.9.13), $J = ds_1(y_1)/dy_1$ and Eq. (3.9.13) becomes the same as (3.8.3). The Jacobian merely generalizes the derivative of the inverse of a single function of one variable to n functions of n variables.

Example 3.9.9

The Joint p.d.f. of the Quotient and the Product of Two Random Variables. Suppose that two random variables X_1 and X_2 have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x_1, x_2) = \begin{cases} 4x_1x_2 & \text{for } 0 < x_1 < 1 \text{ and } 0 < x_2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall determine the joint p.d.f. of two new random variables Y_1 and Y_2, which are defined by the relations

$$Y_1 = \frac{X_1}{X_2} \text{ and } Y_2 = X_1X_2.$$

In the notation of Theorem 3.9.5, we would say that $Y_1 = r_1(X_1, X_2)$ and $Y_2 = r_2(X_1, X_2)$, where

$$r_1(x_1, x_2) = \frac{x_1}{x_2} \text{ and } r_2(x_1, x_2) = x_1x_2. \tag{3.9.14}$$

The inverse of the transformation in Eq. (3.9.14) is found by solving the equations $y_1 = r_1(x_1, x_2)$ and $y_2 = r_2(x_1, x_2)$ for x_1 and x_2 in terms of y_1 and y_2. The result is

$$\begin{aligned} x_1 &= s_1(y_1, y_2) = (y_1y_2)^{1/2}, \\ x_2 &= s_2(y_1, y_2) = \left(\frac{y_2}{y_1}\right)^{1/2}. \end{aligned} \tag{3.9.15}$$

Let S denote the set of points (x_1, x_2) such that $0 < x_1 < 1$ and $0 < x_2 < 1$, so that $\Pr[(X_1, X_2) \in S] = 1$. Let T be the set of (y_1, y_2) pairs such that $(y_1, y_2) \in T$ if and only if $(s_1(y_1, y_2), s_2(y_1, y_2)) \in S$. Then $\Pr[(Y_1, Y_2) \in T] = 1$. The transformation defined by the equations in (3.9.14) or, equivalently, by the equations in (3.9.15) specifies a one-to-one relation between the points in S and the points in T.

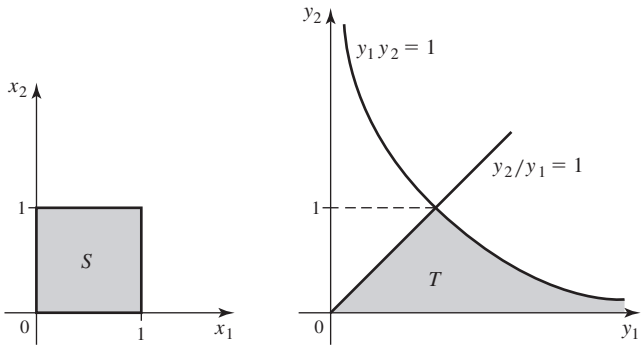

Figure 3.27 The sets S and T in Example 3.9.9.

We shall now show how to find the set T. We know that $(x_1, x_2) \in S$ if and only if the following inequalities hold:

$$x_1 > 0, \quad x_1 < 1, \quad x_2 > 0, \text{ and } x_2 < 1. \tag{3.9.16}$$

We can substitute the formulas for x_1 and x_2 in terms of y_1 and y_2 from Eq. (3.9.15) into the inequalities in (3.9.16) to obtain

$$(y_1y_2)^{1/2} > 0, \quad (y_1y_2)^{1/2} < 1, \quad \left(\frac{y_2}{y_1}\right)^{1/2} > 0,$$
$$\text{and } \left(\frac{y_2}{y_1}\right)^{1/2} < 1. \tag{3.9.17}$$

The first inequality transforms to ($y_1 > 0$ and $y_2 > 0$) or ($y_1 < 0$ and $y_2 < 0$). However, since $y_1 = x_1/x_2$, we cannot have $y_1 < 0$, so we get only $y_1 > 0$ and $y_2 > 0$. The third inequality in (3.9.17) transforms to the same thing. The second inequality in (3.9.17) becomes $y_2 < 1/y_1$. The fourth inequality becomes $y_2 < y_1$. The region T where (y_1, y_2) satisfy these new inequalities is shown in the right panel of Fig. 3.27 with the set S in the left panel.

For the functions in (3.9.15),

$$\frac{\partial s_1}{\partial y_1} = \frac{1}{2}\left(\frac{y_2}{y_1}\right)^{1/2}, \quad \frac{\partial s_1}{\partial y_2} = \frac{1}{2}\left(\frac{y_1}{y_2}\right)^{1/2},$$
$$\frac{\partial s_2}{\partial y_1} = -\frac{1}{2}\left(\frac{y_2}{y_1^3}\right)^{1/2}, \quad \frac{\partial s_2}{\partial y_2} = \frac{1}{2}\left(\frac{1}{y_1y_2}\right)^{1/2}.$$

Hence,

$$J = \det\begin{bmatrix} \frac{1}{2}\left(\frac{y_2}{y_1}\right)^{1/2} & \frac{1}{2}\left(\frac{y_1}{y_2}\right)^{1/2} \\ -\frac{1}{2}\left(\frac{y_2}{y_1^3}\right)^{1/2} & \frac{1}{2}\left(\frac{1}{y_1y_2}\right)^{1/2} \end{bmatrix} = \frac{1}{2y_1}.$$

Since $y_1 > 0$ throughout the set T, $|J| = 1/(2y_1)$.

The joint p.d.f. $g(y_1, y_2)$ can now be obtained directly from Eq. (3.9.13) in the following way: In the expression for $f(x_1, x_2)$, replace x_1 with $(y_1y_2)^{1/2}$, replace x_2

with $(y_2/y_1)^{1/2}$, and multiply the result by $|J| = 1/(2y_1)$. Therefore,

$$g(y_1, y_2) = \begin{cases} 2\left(\frac{y_2}{y_1}\right) & \text{for } (y_1, y_2) \in T, \\ 0 & \text{otherwise.} \end{cases}$$ ◀

Example 3.9.10 Service Time in a Queue. Let X be the time that the server in a single-server queue will spend on a particular customer, and let Y be the rate at which the server can operate. A popular model for the conditional distribution of X given $Y = y$ is to say that the conditional p.d.f. of X given $Y = y$ is

$$g_1(x|y) = \begin{cases} ye^{-xy} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let Y have the p.d.f. $f_2(y)$. The joint p.d.f. of (X, Y) is then $g_1(x|y) f_2(y)$. Because $1/Y$ can be interpreted as the average service time, $Z = XY$ measures how quickly, compared to average, that the customer is served. For example, $Z = 1$ corresponds to an average service time, while $Z > 1$ means that this customer took longer than average, and $Z < 1$ means that this customer was served more quickly than the average customer. If we want the distribution of Z, we could compute the joint p.d.f. of (Z, Y) directly using the methods just illustrated. We could then integrate the joint p.d.f. over y to obtain the marginal p.d.f. of Z. However, it is simpler to transform the conditional distribution of X given $Y = y$ into the conditional distribution of Z given $Y = y$, since conditioning on $Y = y$ allows us to treat Y as the constant y. Because $X = Z/Y$, the inverse transformation is $x = s(z)$, where $s(z) = z/y$. The derivative of this is $1/y$, and the conditional p.d.f. of Z given $Y = y$ is

$$h_1(z|y) = \frac{1}{y} g_1\left(\frac{z}{y} \middle| y\right).$$

Because Y is a rate, $Y \geq 0$ and $X = Z/Y > 0$ if and only if $Z > 0$. So,

$$h_1(z|y) = \begin{cases} e^{-z} & \text{for } z > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9.18}$$

Notice that h_1 does not depend on y, so Z is independent of Y and h_1 is the marginal p.d.f. of Z. The reader can verify all of this in Exercise 17. ◀

Note: Removing Dependence. The formula $Z = XY$ in Example 3.9.10 makes it look as if Z should depend on Y. In reality, however, multiplying X by Y removes the dependence that X already has on Y and makes the result independent of Y. This type of transformation that removes the dependence of one random variable on another is a very powerful technique for finding marginal distributions of transformations of random variables.

In Example 3.9.10, we mentioned that there was another, more straightforward but more tedious, way to compute the distribution of Z. That method, which is useful in many settings, is to transform (X, Y) into (Z, W) for some uninteresting random variable W and then integrate w out of the joint p.d.f. All that matters in the choice of W is that the transformation be one-to-one with differentiable inverse and that the calculations are feasible. Here is a specific example.

Example 3.9.11 One Function of Two Variables. In Example 3.9.9, suppose that we were interested only in the quotient $Y_1 = X_1/X_2$ rather than both the quotient and the product $Y_2 = X_1X_2$. Since we already have the joint p.d.f. of (Y_1, Y_2), we will merely integrate y_2 out rather than start from scratch. For each value of $y_1 > 0$, we need to look at the set T in Fig. 3.27 and find the interval of y_2 values to integrate over. For $0 < y_1 < 1$,

we integrate over $0 < y_2 < y_1$. For $y_1 > 1$, we integrate over $0 < y_2 < 1/y_1$. (For $y_1 = 1$ both intervals are the same.) So, the marginal p.d.f. of Y_1 is

$$g_1(y_1) = \begin{cases} \int_0^{y_1} 2\left(\frac{y_2}{y_1}\right) dy_2 & \text{for } 0 < y_1 < 1, \\ \int_0^{1/y_1} 2\left(\frac{y_2}{y_1}\right) dy_2 & \text{for } y_1 > 1, \end{cases}$$
$$= \begin{cases} y_1 & \text{for } 0 < y_1 < 1, \\ \frac{1}{y_1^3} & \text{for } y_1 > 1. \end{cases}$$

There are other transformations that would have made the calculation of g_1 simpler if that had been all we wanted. See Exercise 21 for an example. ◀

Theorem 3.9.6 **Linear Transformations.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ have a continuous joint distribution for which the joint p.d.f. is f. Define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ by

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}, \tag{3.9.19}$$

where $\boldsymbol{A}$ is a nonsingular $n \times n$ matrix. Then $\boldsymbol{Y}$ has a continuous joint distribution with p.d.f.

$$g(\boldsymbol{y}) = \frac{1}{|\det \boldsymbol{A}|} f(\boldsymbol{A}^{-1}\boldsymbol{y}) \quad \text{for } \boldsymbol{y} \in R^n, \tag{3.9.20}$$

where $\boldsymbol{A}^{-1}$ is the inverse of $\boldsymbol{A}$.

Proof Each Y_i is a linear combination of $X_1, \ldots, X_n$. Because $\boldsymbol{A}$ is nonsingular, the transformation in Eq. (3.9.19) is a one-to-one transformation of the entire space R^n onto itself. At every point $\boldsymbol{y} \in R^n$, the inverse transformation can be represented by the equation

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}. \tag{3.9.21}$$

The Jacobian J of the transformation that is defined by Eq. (3.9.21) is simply $J = \det \boldsymbol{A}^{-1}$. Also, it is known from the theory of determinants that

$$\det \boldsymbol{A}^{-1} = \frac{1}{\det \boldsymbol{A}}.$$

Therefore, at every point $\boldsymbol{y} \in R^n$, the joint p.d.f. $g(\boldsymbol{y})$ can be evaluated in the following way, according to Theorem 3.9.5: First, for $i = 1, \ldots, n$, the component x_i in $f(x_1, \ldots, x_n)$ is replaced with the ith component of the vector $\boldsymbol{A}^{-1}\boldsymbol{y}$. Then, the result is divided by $|\det \boldsymbol{A}|$. This produces Eq. (3.9.20). ■

Summary

We extended the construction of the distribution of a function of a random variable to the case of several functions of several random variables. If one only wants the distribution of one function r_1 of n random variables, the usual way to find this is to first find $n - 1$ additional functions $r_2, \ldots, r_n$ so that the n functions together compose a one-to-one transformation. Then find the joint p.d.f. of the n functions and finally find the marginal p.d.f. of the first function by integrating out the extra $n - 1$ variables. The method is illustrated for the cases of the sum and the range of several random variables.

Exercises

1. Suppose that X_1 and X_2 are i.i.d. random variables and that each of them has the uniform distribution on the interval [0, 1]. Find the p.d.f. of $Y = X_1 + X_2$.

2. For the conditions of Exercise 1, find the p.d.f. of the average $(X_1 + X_2)/2$.

3. Suppose that three random variables X_1, X_2, and X_3 have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x_1, x_2, x_3) = \begin{cases} 8x_1x_2x_3 & \text{for } 0 < x_i < 1 \ (i = 1, 2, 3), \\ 0 & \text{otherwise.} \end{cases}$$

Suppose also that $Y_1 = X_1$, $Y_2 = X_1X_2$, and $Y_3 = X_1X_2X_3$. Find the joint p.d.f. of Y_1, Y_2, and Y_3.

4. Suppose that X_1 and X_2 have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x_1, x_2) = \begin{cases} x_1 + x_2 & \text{for } 0 < x_1 < 1 \text{ and } 0 < x_2 < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the p.d.f. of $Y = X_1X_2$.

5. Suppose that the joint p.d.f. of X_1 and X_2 is as given in Exercise 4. Find the p.d.f. of $Z = X_1/X_2$.

6. Let X and Y be random variables for which the joint p.d.f. is as follows:

$$f(x, y) = \begin{cases} 2(x + y) & \text{for } 0 \le x \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the p.d.f. of $Z = X + Y$.

7. Suppose that X_1 and X_2 are i.i.d. random variables and that the p.d.f. of each of them is as follows:

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

Find the p.d.f. of $Y = X_1 - X_2$.

8. Suppose that $X_1, \ldots, X_n$ form a random sample of size n from the uniform distribution on the interval [0, 1] and that $Y_n = \max\{X_1, \ldots, X_n\}$. Find the smallest value of n such that

$$\Pr\{Y_n \ge 0.99\} \ge 0.95.$$

9. Suppose that the n variables $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval [0, 1] and that the random variables Y_1 and Y_n are defined as in Eq. (3.9.8). Determine the value of $\Pr(Y_1 \le 0.1$ and $Y_n \le 0.8)$.

10. For the conditions of Exercise 9, determine the value of $\Pr(Y_1 \le 0.1$ and $Y_n \ge 0.8)$.

11. For the conditions of Exercise 9, determine the probability that the interval from Y_1 to Y_n will not contain the point 1/3.

12. Let W denote the range of a random sample of n observations from the uniform distribution on the interval [0, 1]. Determine the value of $\Pr(W > 0.9)$.

13. Determine the p.d.f. of the range of a random sample of n observations from the uniform distribution on the interval $[-3, 5]$.

14. Suppose that $X_1, \ldots, X_n$ form a random sample of n observations from the uniform distribution on the interval [0, 1], and let Y denote the second largest of the observations. Determine the p.d.f. of Y. *Hint:* First determine the c.d.f. G of Y by noting that

$$\begin{aligned} G(y) &= \Pr(Y \le y) \\ &= \Pr(\text{At least } n - 1 \text{ observations} \le y). \end{aligned}$$

15. Show that if $X_1, X_2, \ldots, X_n$ are independent random variables and if $Y_1 = r_1(X_1), Y_2 = r_2(X_2), \ldots, Y_n = r_n(X_n)$, then $Y_1, Y_2, \ldots, Y_n$ are also independent random variables.

16. Suppose that $X_1, X_2, \ldots, X_5$ are five random variables for which the joint p.d.f. can be factored in the following form for all points $(x_1, x_2, \ldots, x_5) \in R^5$:

$$f(x_1, x_2, \ldots, x_5) = g(x_1, x_2)h(x_3, x_4, x_5),$$

where g and h are certain nonnegative functions. Show that if $Y_1 = r_1(X_1, X_2)$ and $Y_2 = r_2(X_3, X_4, X_5)$, then the random variables Y_1 and Y_2 are independent.

17. In Example 3.9.10, use the Jacobian method (3.9.13) to verify that Y and Z are independent and that Eq. (3.9.18) is the marginal p.d.f. of Z.

18. Let the conditional p.d.f. of X given Y be $g_1(x|y) = 3x^2/y^3$ for $0 < x < y$ and 0 otherwise. Let the marginal p.d.f. of Y be $f_2(y)$, where $f_2(y) = 0$ for $y \le 0$ but is otherwise unspecified. Let $Z = X/Y$. Prove that Z and Y are independent and find the marginal p.d.f. of Z.

19. Let X_1 and X_2 be as in Exercise 7. Find the p.d.f. of $Y = X_1 + X_2$.

20. If $a_2 = 0$ in Theorem 3.9.4, show that Eq. (3.9.2) becomes the same as Eq. (3.8.1) with $a = a_1$ and $f = f_1$.

21. In Examples 3.9.9 and 3.9.11, find the marginal p.d.f. of $Z_1 = X_1/X_2$ by first transforming to Z_1 and $Z_2 = X_1$ and then integrating z_2 out of the joint p.d.f.

⋆ 3.10 Markov Chains

A popular model for systems that change over time in a random manner is the Markov chain model. A Markov chain is a sequence of random variables, one for each time. At each time, the corresponding random variable gives the state of the system. Also, the conditional distribution of each future state given the past states and the present state depends only on the present state.

Stochastic Processes

Example 3.10.1

Occupied Telephone Lines. Suppose that a certain business office has five telephone lines and that any number of these lines may be in use at any given time. During a certain period of time, the telephone lines are observed at regular intervals of 2 minutes and the number of lines that are being used at each time is noted. Let X_1 denote the number of lines that are being used when the lines are first observed at the beginning of the period; let X_2 denote the number of lines that are being used when they are observed the second time, 2 minutes later; and in general, for $n = 1, 2, \ldots$, let X_n denote the number of lines that are being used when they are observed for the nth time. ◀

Definition 3.10.1

Stochastic Process. A sequence of random variables $X_1, X_2, \ldots$ is called a *stochastic process* or *random process* with *discrete time parameter*. The first random variable X_1 is called the *initial state* of the process; and for $n = 2, 3, \ldots$, the random variable X_n is called the *state of the process at time n*.

In Example 3.10.1, the state of the process at any time is the number of lines being used at that time. Therefore, each state must be an integer between 0 and 5. Each of the random variables in a stochastic process has a marginal distribution, and the entire process has a joint distribution. For convenience, in this text, we will discuss only joint distributions for finitely many of $X_1, X_2, \ldots$ at a time. The meaning of the phrase "discrete time parameter" is that the process, such as the numbers of occupied phone lines, is observed only at discrete or separated points in time, rather than continuously in time. In Sec. 5.4, we will introduce a different stochastic process (called the Poisson process) with a continuous time parameter.

In a stochastic process with a discrete time parameter, the state of the process varies in a random manner from time to time. To describe a complete probability model for a particular process, it is necessary to specify the distribution for the initial state X_1 and also to specify for each $n = 1, 2, \ldots$ the conditional distribution of the subsequent state X_{n+1} given $X_1, \ldots, X_n$. These conditional distributions are equivalent to the collection of conditional c.d.f.'s of the following form:

$$\Pr(X_{n+1} \le b | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

Markov Chains

A Markov chain is a special type of stochastic process, defined in terms of the conditional distributions of future states given the present and past states.

Definition 3.10.2

Markov Chain. A stochastic process with discrete time parameter is a *Markov chain* if, for each time n, the conditional distributions of all X_{n+j} for $j \ge 1$ given $X_1, \ldots, X_n$ depend only on X_n and not on the earlier states $X_1, \ldots, X_{n-1}$. In symbols, for

$n = 1, 2, \ldots$ and for each b and each possible sequence of states $x_1, x_2, \ldots, x_n$,

$$\Pr(X_{n+1} \le b | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_{n+1} \le b | X_n = x_n).$$

A Markov chain is called *finite* if there are only finitely many possible states.

In the remainder of this section, we shall consider only finite Markov chains. This assumption could be relaxed at the cost of more complicated theory and calculation. For convenience, we shall reserve the symbol k to stand for the number of possible states of a general finite Markov chain for the remainder of the section. It will also be convenient, when discussing a general finite Markov chain, to name the k states using the integers $1, \ldots, k$. That is, for each n and j, $X_n = j$ will mean that the chain is in state j at time n. In specific examples, it may prove more convenient to label the states in a more informative fashion. For example, if the states are the numbers of phone lines in use at given times (as in the example that introduced this section), we would label the states $0, \ldots, 5$ even though $k = 6$.

The following result follows from the multiplication rule for conditional probabilities, Theorem 2.1.2.

Theorem 3.10.1

For a finite Markov chain, the joint p.f. for the first n states equals

$$\begin{aligned}&\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\\ &\quad = \Pr(X_1 = x_1) \Pr(X_2 = x_2 | X_1 = x_1) \Pr(X_3 = x_3 | X_2 = x_2) \cdots\\ &\qquad \Pr(X_n = x_n | X_{n-1} = x_{n-1}).\end{aligned} \tag{3.10.1}$$

Also, for each n and each $m > 0$,

$$\begin{aligned}&\Pr(X_{n+1} = x_{n+1}, X_{n+2} = x_{n+2}, \ldots, X_{n+m} = x_{n+m} | X_n = x_n)\\ &\quad = \Pr(X_{n+1} = x_{n+1} | X_n = x_n) \Pr(X_{n+2} = x_{n+2} | X_{n+1} = x_{n+1})\\ &\qquad \cdots \Pr(X_{n+m} = x_{n+m} | X_{n+m-1} = x_{n+m-1}).\end{aligned} \tag{3.10.2}$$

■

Eq. (3.10.1) is a discrete version of a generalization of conditioning in sequence that was illustrated in Example 3.7.18 with continuous random variables. Eq. (3.10.2) is a conditional version of (3.10.1) shifted forward in time.

Example 3.10.2

Shopping for Toothpaste. In Exercise 4 in Sec. 2.1, we considered a shopper who chooses between two brands of toothpaste on several occasions. Let $X_i = 1$ if the shopper chooses brand A on the ith purchase, and let $X_i = 2$ if the shopper chooses brand B on the ith purchase. Then the sequence of states $X_1, X_2, \ldots$ is a stochastic process with two possible states at each time. The probabilities of purchase were specified by saying that the shopper will choose the same brand as on the previous purchase with probability 1/3 and will switch with probability 2/3. Since this happens regardless of purchases that are older than the previous one, we see that this stochastic process is a Markov chain with

$$\Pr(X_{n+1} = 1 | X_n = 1) = \frac{1}{3}, \ \Pr(X_{n+1} = 2 | X_n = 1) = \frac{2}{3},$$
$$\Pr(X_{n+1} = 1 | X_n = 2) = \frac{2}{3}, \ \Pr(X_{n+1} = 2 | X_n = 2) = \frac{1}{3}.$$

◀

Exammple 3.10.2 has an additional feature that puts it in a special class of Markov chains. The probability of moving from one state at time n to another state at time $n + 1$ does not depend on n.

Definition 3.10.3 Transition Distributions/Stationary Transition Distributions. Consider a finite Markov chain with k possible states. The conditional distributions of the state at time $n+1$ given the state at time n, that is, $\Pr(X_{n+1} = j | X_n = i)$ for $i, j = 1, \ldots, k$ and $n = 1, 2, \ldots$, are called the *transition distributions* of the Markov chain. If the transition distribution is the same for every time n ($n = 1, 2, \ldots$), then the Markov chain has *stationary transition distributions*.

When a Markov chain with k possible states has stationary transition distributions, there exist probabilities p_{ij} for $i, j = 1, \ldots, k$ such that, for all n,

$$\Pr(X_{n+1} = j | X_n = i) = p_{ij} \quad \text{for } n = 1, 2, \ldots. \tag{3.10.3}$$

The Markov chain in Example 3.10.2 has stationary transition distributions. For example, $p_{11} = 1/3$.

In the language of multivariate distributions, when a Markov chain has stationary transition distributions, specified by (3.10.3), we can write the conditional p.f. of X_{n+1} given X_n as

$$g(j|i) = p_{ij}, \tag{3.10.4}$$

for all n, i, j.

Example 3.10.3 Occupied Telephone Lines. To illustrate the application of these concepts, we shall consider again the example involving the office with five telephone lines. In order for this stochastic process to be a Markov chain, the specified distribution for the number of lines that may be in use at each time must depend only on the number of lines that were in use when the process was observed most recently 2 minutes earlier and must not depend on any other observed values previously obtained. For example, if three lines were in use at time n, then the distribution for time $n+1$ must be the same regardless of whether 0, 1, 2, 3, 4, or 5 lines were in use at time $n-1$. In reality, however, the observation at time $n-1$ might provide some information in regard to the length of time for which each of the three lines in use at time n had been occupied, and this information might be helpful in determining the distribution for time $n+1$. Nevertheless, we shall suppose now that this process is a Markov chain. If this Markov chain is to have stationary transition distributions, it must be true that the rates at which incoming and outgoing telephone calls are made and the average duration of these telephone calls do not change during the entire period covered by the process. This requirement means that the overall period cannot include busy times when more calls are expected or quiet times when fewer calls are expected. For example, if only one line is in use at a particular observation time, regardless of when this time occurs during the entire period covered by the process, then there must be a specific probability p_{1j} that exactly j lines will be in use 2 minutes later. ◀

The Transition Matrix

Example 3.10.4 Shopping for Toothpaste. The notation for stationary transition distributions, p_{ij}, suggests that they could be arranged in a matrix. The transition probabilities for Example 3.10.2 can be arranged into the following matrix:

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}. \quad \blacktriangleleft$$

Every finite Markov chain with stationary transition distributions has a matrix like the one constructed in Example 3.10.4.

Definition 3.10.4 Transition Matrix. Consider a finite Markov chain with stationary transition distributions given by $p_{ij} = \Pr(X_{n+1} = j | X_n = i)$ for all n, i, j. The *transition matrix* of the Markov chain is defined to be the $k \times k$ matrix $\boldsymbol{P}$ with elements p_{ij}. That is,

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & \cdots & p_{1k} \\ p_{21} & \cdots & p_{2k} \\ \vdots & \ddots & \vdots \\ p_{k1} & \cdots & p_{kk} \end{bmatrix}. \tag{3.10.5}$$

A transition matrix has several properties that are apparent from its defintion. For example, each element is nonnegative because all elements are probabilities. Since each row of a transition matrix is a conditional p.f. for the next state given some value of the current state, we have $\sum_{j=1}^{k} p_{ij} = 1$ for $i = 1, \ldots, k$. Indeed, row i of the transition matrix specifies the conditional p.f. $g(\cdot|i)$ defined in (3.10.4).

Definition 3.10.5 Stochastic Matrix. A square matrix for which all elements are nonnegative and the sum of the elements in each row is 1 is called a *stochastic matrix*.

It is clear that the transition matrix $\boldsymbol{P}$ for every finite Markov chain with stationary transition probabilities must be a stochastic matrix. Conversely, every $k \times k$ stochastic matrix can serve as the transition matrix of a finite Markov chain with k possible states and stationary transition distributions.

Example 3.10.5 A Transition Matrix for the Number of Occupied Telephone Lines. Suppose that in the example involving the office with five telephone lines, the numbers of lines being used at times 1, 2, ... form a Markov chain with stationary transition distributions. This chain has six possible states 0, 1, ..., 5, where i is the state in which exactly i lines are being used at a given time ($i = 0, 1, \ldots, 5$). Suppose that the transition matrix $\boldsymbol{P}$ is as follows:

$$\boldsymbol{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.2 & 0.3 & 0.2 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.2 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.2 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.4 & 0.2 \end{bmatrix} \end{array}. \tag{3.10.6}$$

(a) Assuming that all five lines are in use at a certain observation time, we shall determine the probability that exactly four lines will be in use at the next observation time. (b) Assuming that no lines are in use at a certain time, we shall determine the probability that at least one line will be in use at the next observation time.

(a) This probability is the element in the matrix $\boldsymbol{P}$ in the row corresponding to the state 5 and the column corresponding to the state 4. Its value is seen to be 0.4.

(b) If no lines are in use at a certain time, then the element in the upper left corner of the matrix $\boldsymbol{P}$ gives the probability that no lines will be in use at the next observation time. Its value is seen to be 0.1. Therefore, the probability that at least one line will be in use at the next observation time is $1 - 0.1 = 0.9$. ◀

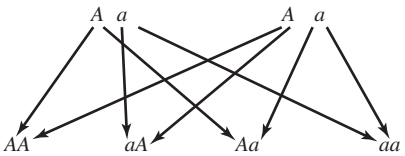

Figure 3.28 The generation following $\{Aa, Aa\}$.

Example 3.10.6

Plant Breeding Experiment. A botanist is studying a certain variety of plant that is monoecious (has male and female organs in separate flowers on a single plant). She begins with two plants I and II and cross-pollinates them by crossing male I with female II and female I with male II to produce two offspring for the next generation. The original plants are destroyed and the process is repeated as soon as the new generation of two plants is mature. Several replications of the study are run simultaneously. The botanist might be interested in the proportion of plants in any generation that have each of several possible genotypes for a particular gene. (See Example 1.6.4 on page 23.) Suppose that the gene has two alleles, A and a. The genotype of an individual will be one of the three combinations AA, Aa, or aa. When a new individual is born, it gets one of the two alleles (with probability 1/2 each) from one of the parents, and it independently gets one of the two alleles from the other parent. The two offspring get their genotypes independently of each other. For example, if the parents have genotypes AA and Aa, then an offspring will get A for sure from the first parent and will get either A or a from the second parent with probability 1/2 each. Let the states of this population be the set of genotypes of the two members of the current population. We will not distinguish the set $\{AA, Aa\}$ from $\{Aa, AA\}$. There are then six states: $\{AA, AA\}$, $\{AA, Aa\}$, $\{AA, aa\}$, $\{Aa, Aa\}$, $\{Aa, aa\}$, and $\{aa, aa\}$. For each state, we can calculate the probability that the next generation will be in each of the six states. For example, if the state is either $\{AA, AA\}$ or $\{aa, aa\}$, the next generation will be in the same state with probability 1. If the state is $\{AA, aa\}$, the next generation will be in state $\{Aa, Aa\}$ with probability 1. The other three states have more complicated transitions.

If the current state is $\{Aa, Aa\}$, then all six states are possible for the next generation. In order to compute the transition distribution, it helps to first compute the probability that a given offspring will have each of the three genotypes. Figure 3.28 illustrates the possible offspring in this state. Each arrow going down in Fig. 3.28 is a possible inheritance of an allele, and each combination of arrows terminating in a genotype has probability 1/4. It follows that the probability of AA and aa are both 1/4, while the probability of Aa is 1/2, because two different combinations of arrows lead to this offspring. In order for the next state to be $\{AA, AA\}$, both offspring must be AA independently, so the probability of this transition is 1/16. The same argument implies that the probability of a transition to $\{aa, aa\}$ is 1/16. A transition to $\{AA, Aa\}$ requires one offspring to be AA (probability 1/4) and the other to be Aa (probabilty 1/2). But the two different genotypes could occur in either order, so the whole probability of such a transition is $2 \times (1/4) \times (1/2) = 1/4$. A similar argument shows that a transition to $\{Aa, aa\}$ also has probability 1/4. A transition to $\{AA, aa\}$ requires one offspring to be AA (probability 1/4) and the other to be aa (probability 1/4). Once again, these can occur in two orders, so the whole probability is $2 \times 1/4 \times 1/4 = 1/8$. By subtraction, the probability of a transition to $\{Aa, Aa\}$ must be $1 - 1/16 - 1/16 - 1/4 - 1/4 - 1/8 = 1/4$. Here is the entire transition matrix, which can be verified in a manner similar to what has just been done:

$$\begin{array}{c} \\ \{AA, AA\} \\ \{AA, Aa\} \\ \{AA, aa\} \\ \{Aa, Aa\} \\ \{Aa, aa\} \\ \{aa, aa\} \end{array} \begin{array}{c} \begin{array}{cccccc} \{AA, AA\} & \{AA, Aa\} & \{AA, aa\} & \{Aa, Aa\} & \{Aa, aa\} & \{aa, aa\} \end{array} \\ \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2500 & 0.5000 & 0.0000 & 0.2500 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0625 & 0.2500 & 0.1250 & 0.2500 & 0.2500 & 0.0625 \\ 0.0000 & 0.0000 & 0.0000 & 0.2500 & 0.5000 & 0.2500 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{bmatrix} \end{array}.$$

◀

The Transition Matrix for Several Steps

Example 3.10.7

Single Server Queue. A manager usually checks the server at her store every 5 minutes to see whether the server is busy or not. She models the state of the server (1 = busy or 2 = not busy) as a Markov chain with two possible states and stationary transition distributions given by the following matrix:

$$\boldsymbol{P} = \begin{array}{c} \\ \text{Busy} \\ \text{Not busy} \end{array} \begin{array}{c} \begin{array}{cc} \text{Busy} & \text{Not busy} \end{array} \\ \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \end{bmatrix} \end{array}.$$

The manager realizes that, later in the day, she will have to be away for 10 minutes and will miss one server check. She wants to compute the conditional distribution of the state two time periods in the future given each of the possible states. She reasons as follows: If $X_n = 1$ for example, then the state will have to be either 1 or 2 at time $n + 1$ even though she does not care now about the state at time $n + 1$. But, if she computes the joint conditional distribution of X_{n+1} and X_{n+2} given $X_n = 1$, she can sum over the possible values of X_{n+1} to get the conditional distribution of X_{n+2} given $X_n = 1$. In symbols,

$$\begin{aligned} \Pr(X_{n+2} = 1 | X_n = 1) = {} & \Pr(X_{n+1} = 1,\, X_{n+2} = 1 | X_n = 1) \\ & + \Pr(X_{n+1} = 2,\, X_{n+2} = 1 | X_n = 1). \end{aligned}$$

By the second part of Theorem 3.10.1,

$$\begin{aligned} \Pr(X_{n+1} = 1,\, X_{n+2} = 1 | X_n = 1) &= \Pr(X_{n+1} = 1 | X_n = 1) \Pr(X_{n+2} = 1 | X_{n+1} = 1) \\ &= 0.9 \times 0.9 = 0.81. \end{aligned}$$

Similarly,

$$\begin{aligned} \Pr(X_{n+1} = 2,\, X_{n+2} = 1 | X_n = 1) &= \Pr(X_{n+1} = 2 | X_n = 1) \Pr(X_{n+2} = 1 | X_{n+1} = 2) \\ &= 0.1 \times 0.6 = 0.06. \end{aligned}$$

It follows that $\Pr(X_{n+2} = 1 | X_n = 1) = 0.81 + 0.06 = 0.87$, and hence $\Pr(X_{n+2} = 2 | X_n = 1) = 1 - 0.87 = 0.13$. By similar reasoning, if $X_n = 2$,

$$\Pr(X_{n+2} = 1 | X_n = 2) = 0.6 \times 0.9 + 0.4 \times 0.6 = 0.78,$$

and $\Pr(X_{n+2} = 2 | X_n = 2) = 1 - 0.78 = 0.22$. ◀

Generalizing the calculations in Example 3.10.7 to three or more transitions might seem tedious. However, if one examines the calculations carefully, one sees a pattern

that will allow a compact calculation of transition distributions for several steps. Consider a general Markov chain with k possible states $1, \ldots, k$ and the transition matrix $\boldsymbol{P}$ given by Eq. (3.10.5). Assuming that the chain is in state i at a given time n, we shall now determine the probability that the chain will be in state j at time $n+2$. In other words, we shall determine the conditional probability of $X_{n+2} = j$ given $X_n = i$. The notation for this probability is $p_{ij}^{(2)}$.

We argue as the manager did in Example 3.10.7. Let r denote the value of X_{n+1} that is not of primary interest but is helpful to the calculation. Then

$$\begin{aligned} p_{ij}^{(2)} &= \Pr(X_{n+2} = j | X_n = i) \\ &= \sum_{r=1}^{k} \Pr(X_{n+1} = r \quad \text{and} \quad X_{n+2} = j | X_n = i) \\ &= \sum_{r=1}^{k} \Pr(X_{n+1} = r | X_n = i) \Pr(X_{n+2} = j | X_{n+1} = r, X_n = i) \\ &= \sum_{r=1}^{k} \Pr(X_{n+1} = r | X_n = i) \Pr(X_{n+2} = j | X_{n+1} = r) \\ &= \sum_{r=1}^{k} p_{ir} p_{rj}, \end{aligned}$$

where the third equality follows from Theorem 2.1.3 and the fourth equality follows from the definition of a Markov chain.

The value of $p_{ij}^{(2)}$ can be determined in the following manner: If the transition matrix $\boldsymbol{P}$ is squared, that is, if the matrix $\boldsymbol{P}^2 = \boldsymbol{PP}$ is constructed, then the element in the ith row and the jth column of the matrix $\boldsymbol{P}^2$ will be $\sum_{r=1}^{k} p_{ir} p_{rj}$. Therefore, $p_{ij}^{(2)}$ will be the element in the ith row and the jth column of $\boldsymbol{P}^2$.

By a similar argument, the probability that the chain will move from the state i to the state j in three steps, or $p_{ij}^{(3)} = \Pr(X_{n+3} = j | X_n = i)$, can be found by constructing the matrix $\boldsymbol{P}^3 = \boldsymbol{P}^2\boldsymbol{P}$. Then the probability $p_{ij}^{(3)}$ will be the element in the ith row and the jth column of the matrix $\boldsymbol{P}^3$.

In general, we have the following result.

Theorem 3.10.2 Multiple Step Transitions. Let $\boldsymbol{P}$ be the transition matrix of a finite Markov chain with stationary transition distributions. For each $m = 2, 3, \ldots$, the mth power $\boldsymbol{P}^m$ of the matrix $\boldsymbol{P}$ has in row i and column j the probability $p_{ij}^{(m)}$ that the chain will move from state i to state j in m steps. ■

Definition 3.10.6 Multiple Step Transition Matrix. Under the conditions of Theorem 3.10.2, the matrix $\boldsymbol{P}^m$ is called the *m-step transition matrix* of the Markov chain.

In summary, the ith row of the m-step transition matrix gives the conditional distribution of X_{n+m} given $X_n = i$ for all $i = 1, \ldots, k$ and all $n, m = 1, 2, \ldots$.

Example 3.10.8 The Two-Step and Three-Step Transition Matrices for the Number of Occupied Telephone Lines. Consider again the transition matrix $\boldsymbol{P}$ given by Eq. (3.10.6) for the Markov chain based on five telephone lines. We shall assume first that i lines are in use at a

certain time, and we shall determine the probability that exactly j lines will be in use two time periods later.

If we multiply the matrix $\boldsymbol{P}$ by itself, we obtain the following two-step transition matrix:

$$\boldsymbol{P}^2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.14 & 0.23 & 0.20 & 0.15 & 0.16 & 0.12 \\ 0.13 & 0.24 & 0.20 & 0.15 & 0.16 & 0.12 \\ 0.12 & 0.20 & 0.21 & 0.18 & 0.17 & 0.12 \\ 0.11 & 0.17 & 0.19 & 0.20 & 0.20 & 0.13 \\ 0.11 & 0.16 & 0.16 & 0.18 & 0.24 & 0.15 \\ 0.11 & 0.16 & 0.15 & 0.17 & 0.25 & 0.16 \end{bmatrix} \end{array}. \tag{3.10.7}$$

From this matrix we can find any two-step transition probability for the chain, such as the following:

i. If two lines are in use at a certain time, then the probability that four lines will be in use two time periods later is 0.17.

ii. If three lines are in use at a certain time, then the probability that three lines will again be in use two time periods later is 0.20.

We shall now assume that i lines are in use at a certain time, and we shall determine the probability that exactly j lines will be in use three time periods later.

If we construct the matrix $\boldsymbol{P}^3 = \boldsymbol{P}^2\boldsymbol{P}$, we obtain the following three-step transition matrix:

$$\boldsymbol{P}^3 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.123 & 0.208 & 0.192 & 0.166 & 0.183 & 0.128 \\ 0.124 & 0.207 & 0.192 & 0.166 & 0.183 & 0.128 \\ 0.120 & 0.197 & 0.192 & 0.174 & 0.188 & 0.129 \\ 0.117 & 0.186 & 0.186 & 0.179 & 0.199 & 0.133 \\ 0.116 & 0.181 & 0.177 & 0.176 & 0.211 & 0.139 \\ 0.116 & 0.180 & 0.174 & 0.174 & 0.215 & 0.141 \end{bmatrix} \end{array}. \tag{3.10.8}$$

From this matrix we can find any three-step transition probability for the chain, such as the following:

i. If all five lines are in use at a certain time, then the probability that no lines will be in use three time periods later is 0.116.

ii. If one line is in use at a certain time, then the probability that exactly one line will again be in use three time periods later is 0.207. ◀

Example 3.10.9 Plant Breeding Experiment. In Example 3.10.6, the transition matrix has many zeros, since many of the transitions will not occur. However, if we are willing to wait two steps, we will find that the only transitions that cannot occur in two steps are those from the first state to anything else and those from the last state to anything else.

Here is the two-step transition matrix:

$$\begin{array}{c} \\ \{AA, AA\} \\ \{AA, Aa\} \\ \{AA, aa\} \\ \{Aa, Aa\} \\ \{Aa, aa\} \\ \{aa, aa\} \end{array} \begin{array}{c} \begin{array}{cccccc} \{AA, AA\} & \{AA, Aa\} & \{AA, aa\} & \{Aa, Aa\} & \{Aa, aa\} & \{aa, aa\} \end{array} \\ \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3906 & 0.3125 & 0.0313 & 0.1875 & 0.0625 & 0.0156 \\ 0.0625 & 0.2500 & 0.1250 & 0.2500 & 0.2500 & 0.0625 \\ 0.1406 & 0.1875 & 0.0313 & 0.3125 & 0.1875 & 0.1406 \\ 0.0156 & 0.0625 & 0.0313 & 0.1875 & 0.3125 & 0.3906 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{bmatrix} \end{array}.$$

Indeed, if we look at the three-step or the four-step or the general m-step transition matrix, the first and last rows will always be the same. ◀

The first and last states in Example 3.10.9 have the property that, once the chain gets into one of those states, it can't get out. Such states occur in many Markov chains and have a special name.

Definition 3.10.7 Absorbing State. In a Markov chain, if $p_{ii} = 1$ for some state i, then that state is called an *absorbing state*.

In Example 3.10.9, there is positive probability of getting into each absorbing state in two steps no matter where the chain starts. Hence, the probability is 1 that the chain will eventually be absorbed into one of the absorbing states if it is allowed to run long enough.

The Initial Distribution

Example 3.10.10 Single Server Queue. The manager in Example 3.10.7 enters the store thinking that the probability is 0.3 that the server will be busy the first time that she checks. Hence, the probability is 0.7 that the server will be not busy. These values specify the marginal distribution of the state at time 1, X_1. We can represent this distribution by the vector $\boldsymbol{v} = (0.3, 0.7)$ that gives the probabilities of the two states at time 1 in the same order that they appear in the transition matrix. ◀

The vector giving the marginal distribution of X_1 in Example 3.10.10 has a special name.

Definition 3.10.8 Probability Vector/Initial Distribution. A vector consisting of nonnegative numbers that add to 1 is called a *probability vector*. A probability vector whose coordinates specify the probabilities that a Markov chain will be in each of its states at time 1 is called the *initial distribution* of the chain or the *intial probability vector*.

For Example 3.10.2, the initial distribution was given in Exercise 4 in Sec. 2.1 as $\boldsymbol{v} = (0.5, 0.5)$.

The initial distribution and the transition matrix together determine the entire joint distribution of the Markov chain. Indeed, Theorem 3.10.1 shows how to construct the joint distribution of the chain from the initial probability vector and the transition matrix. Letting $\boldsymbol{v} = (v_1, \ldots, v_k)$ denote the initial distribution, Eq. (3.10.1) can be rewritten as

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = v_{x_1} p_{x_1 x_2} \cdots p_{x_{n-1} x_n}. \tag{3.10.9}$$

The marginal distributions of states at times later than 1 can be found from the joint distribution.

Theorem 3.10.3 Marginal Distributions at Times Other Than 1. Consider a finite Markov chain with stationary transition distributions having initial distribution $\boldsymbol{v}$ and transition matrix $\boldsymbol{P}$. The marginal distribution of X_n, the state at time n, is given by the probability vector $\boldsymbol{v}\boldsymbol{P}^{n-1}$.

Proof The marginal distribution of X_n can be found from Eq. (3.10.9) by summing over the possible values of $x_1, \ldots, x_{n-1}$. That is,

$$\Pr(X_n = x_n) = \sum_{x_{n-1}=1}^{k} \cdots \sum_{x_2=1}^{k} \sum_{x_1=1}^{k} v_{x_1} p_{x_1 x_2} p_{x_2 x_3} \cdots p_{x_{n-1} x_n}. \tag{3.10.10}$$

The innermost sum in Eq. (3.10.10) for $x_1 = 1, \ldots, k$ involves only the first two factors $v_{x_1} p_{x_1 x_2}$ and produces the x_2 coordinate of $\boldsymbol{v}\boldsymbol{P}$. Similarly, the next innermost sum over $x_2 = 1, \ldots, k$ involves only the x_2 coordinate of $\boldsymbol{v}\boldsymbol{P}$ and $p_{x_2 x_3}$ and produces the x_3 coordinate of $\boldsymbol{v}\boldsymbol{P}\boldsymbol{P} = \boldsymbol{v}\boldsymbol{P}^2$. Proceeding in this way through all $n-1$ summations produces the x_n coordinate of $\boldsymbol{v}\boldsymbol{P}^{n-1}$. ■

Example 3.10.11 Probabilities for the Number of Occupied Telephone Lines. Consider again the office with five telephone lines and the Markov chain for which the transition matrix $\boldsymbol{P}$ is given by Eq. (3.10.6). Suppose that at the beginning of the observation process at time $n = 1$, the probability that no lines will be in use is 0.5, the probability that one line will be in use is 0.3, and the probability that two lines will be in use is 0.2. Then the initial probability vector is $\boldsymbol{v} = (0.5, 0.3, 0.2, 0, 0, 0)$. We shall first determine the distribution of the number of lines in use at time 2, one period later.

By an elementary computation it will be found that

$$\boldsymbol{v}\boldsymbol{P} = (0.13, 0.33, 0.22, 0.12, 0.10, 0.10).$$

Since the first component of this probability vector is 0.13, the probability that no lines will be in use at time 2 is 0.13; since the second component is 0.33, the probability that exactly one line will be in use at time 2 is 0.33; and so on.

Next, we shall determine the distribution of the number of lines that will be in use at time 3.

By use of Eq. (3.10.7), it can be found that

$$\boldsymbol{v}\boldsymbol{P}^2 = (0.133, 0.227, 0.202, 0.156, 0.162, 0.120).$$

Since the first component of this probability vector is 0.133, the probability that no lines will be in use at time 3 is 0.133; since the second component is 0.227, the probability that exactly one line will be in use at time 3 is 0.227; and so on. ◀

Stationary Distributions

Example 3.10.12 A Special Initial Distribution for Telephone Lines. Suppose that the initial distribution for the number of occupied telephone lines is

$$\boldsymbol{v} = (0.119, 0.193, 0.186, 0.173, 0.196, 0.133).$$

It can be shown, by matrix multiplication, that $\boldsymbol{v}\boldsymbol{P} = \boldsymbol{v}$. This means that if $\boldsymbol{v}$ is the initial distribution, then it is also the distribution after one transition. Hence, it will also be the distribution after two or more transitions as well. ◀

Definition 3.10.9 Stationary Distribution. Let $\boldsymbol{P}$ be the transition matrix for a Markov chain. A probability vector $\boldsymbol{v}$ that satisfies $\boldsymbol{vP} = \boldsymbol{v}$ is called a *stationary distribution* for the Markov chain.

The initial distribution in Example 3.10.12 is a stationary distribution for the telephone lines Markov chain. If the chain starts in this distribution, the distribution stays the same at all times. Every finite Markov chain with stationary transition distributions has at least one stationary distribution. Some chains have a unique stationary distribution.

Note: A Stationary Distribution Does Not Mean That the Chain is Not Moving. It is important to note that $\boldsymbol{vP}^n$ gives the probabilities that the chain is in each of its states after n transitions, calculated before the initial state of the chain or any transitions are observed. These are different from the probabilities of being in the various states after observing the initial state or after observing any of the intervening transitions. In addition, a stationary distribution does not imply that the Markov chain is staying put. If a Markov chain starts in a stationary distribution, then for each state i, the probability that the chain is in state i after n transitions is the same as the probability that it is state i at the start. But the Markov chain can still move around from one state to the next at each transition. The one case in which a Markov chain does stay put is after it moves into an absorbing state. A distribution that is concentrated solely on absorbing states will necessarily be stationary because the Markov chain will never move if it starts in such a distribution. In such cases, all of the uncertainty surrounds the initial state, which will also be the state after every transition.

Example 3.10.13 Stationary Distributions for the Plant Breeding Experiment. Consider again the experiment described in Example 3.10.6. The first and sixth states, $\{AA, AA\}$ and $\{aa, aa\}$, respectively, are absorbing states. It is easy to see that every initial distribution of the form $\boldsymbol{v} = (p, 0, 0, 0, 0, 1-p)$ for $0 \le p \le 1$ has the property that $\boldsymbol{vP} = \boldsymbol{v}$. Suppose that the chain is in state 1 with probability p and in state 6 with probability $1-p$ at time 1. Because these two states are absorbing states, the chain will never move and the event $X_1 = 1$ is the same as the event that $X_n = 1$ for all n. Similarly, $X_1 = 6$ is the same as $X_n = 6$. So, thinking ahead to where the chain is likely to be after n transitions, we would also say that it will be in state 1 with probability p and in state 6 with probability $1-p$. ◀

Method for Finding Stationary Distributions We can rewrite the equation $\boldsymbol{vP} = \boldsymbol{v}$ that defines stationary distributions as $\boldsymbol{v}[\boldsymbol{P} - \boldsymbol{I}] = \boldsymbol{0}$, where $\boldsymbol{I}$ is a $k \times k$ identity matrix and $\boldsymbol{0}$ is a k-dimensional vector of all zeros. Unfortunately, this system of equations has lots of solutions even if there is a unique stationary distribution. The reason is that whenever $\boldsymbol{v}$ solves the system, so does $c\boldsymbol{v}$ for all real c (including $c = 0$). Even though the system has k equations for k variables, there is at least one redundant equation. However, there is also one missing equation. We need to require that the solution vector $\boldsymbol{v}$ has coordinates that sum to 1. We can fix both of these problems by replacing one of the equations in the original system by the equation that says that the coordinates of $\boldsymbol{v}$ sum to 1.

To be specific, define the matrix $\boldsymbol{G}$ to be $\boldsymbol{P} - \boldsymbol{I}$ with its last column replaced by a column of all ones. Then, solve the equation

$$\boldsymbol{v}\boldsymbol{G} = (0, \ldots, 0, 1). \tag{3.10.11}$$

If there is a unique stationary distribution, we will find it by solving (3.10.11). In this case, the matrix $\boldsymbol{G}$ will have an inverse $\boldsymbol{G}^{-1}$ that satisfies

$$\boldsymbol{G}\boldsymbol{G}^{-1} = \boldsymbol{G}^{-1}\boldsymbol{G} = \boldsymbol{I}.$$

The solution of (3.10.11) will then be

$$\boldsymbol{v} = (0, \ldots, 0, 1)\boldsymbol{G}^{-1},$$

which is easily seen to be the bottom row of the matrix $\boldsymbol{G}^{-1}$. This was the method used to find the stationary distribution in Example 3.10.12. If the Markov chain has multiple stationary distributions, then the matrix $\boldsymbol{G}$ will be singular, and this method will not find any of the stationary distributions. That is what would happen in Example 3.10.13 if one were to apply the method.

Example 3.10.14 Stationary Distribution for Toothpaste Shopping. Consider the transition matrix $\boldsymbol{P}$ given in Example 3.10.4. We can construct the matrix $\boldsymbol{G}$ as follows:

$$\boldsymbol{P} - \boldsymbol{I} = \begin{bmatrix} -\frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{2}{3} \end{bmatrix}; \quad \text{hence } \boldsymbol{G} = \begin{bmatrix} -\frac{2}{3} & 1 \\ \frac{2}{3} & 1 \end{bmatrix}.$$

The inverse of $\boldsymbol{G}$ is

$$\boldsymbol{G}^{-1} = \begin{bmatrix} -\frac{3}{4} & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

We now see that the stationary distribution is the bottom row of G^{-1}, $\boldsymbol{v} = (1/2, 1/2)$. ◀

There is a special case in which it is known that a unique stationary distribution exists and it has special properties.

Theorem 3.10.4 If there exists m such that every element of $\boldsymbol{P}^m$ is strictly positive, then

- the Markov chain has a unique stationary distribution $\boldsymbol{v}$,
- $\lim_{n\to\infty} \boldsymbol{P}^n$ is a matrix with all rows equal to $\boldsymbol{v}$, and
- no matter with what distribution the Markov chain starts, its distribution after n steps converges to $\boldsymbol{v}$ as $n \to \infty$. ■

We shall not prove Theorem 3.10.4, although some evidence for the second claim can be seen in Eq. (3.10.8), where the six rows of $\boldsymbol{P}^3$ are much more alike than the rows of $\boldsymbol{P}$ and they are very similar to the stationary distribution given in Example 3.10.12. The third claim in Theorem 3.10.4 actually follows easily from the second claim. In Sec. 12.5, we shall introduce a method that makes use of the third claim in Theorem 3.10.4 in order to approximate distributions of random variables when those distributions are difficult to calculate exactly.

The transition matrices in Examples 3.10.2, 3.10.5, and 3.10.7 satisfy the conditions of Theorem 3.10.4. The following example has a unique stationary distribution but does not satisfy the conditions of Theorem 3.10.4.

Example 3.10.15 Alternating Chain. Let the transition matrix for a two-state Markov chain be

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The matrix $\boldsymbol{G}$ is easy to construct and invert, and we find that the unique stationary distribution is $\boldsymbol{v} = (0.5, 0.5)$. However, as m increases, $\boldsymbol{P}^m$ alternates between $\boldsymbol{P}$ and the 2×2 identity matrix. It does not converge and never does it have all elements strictly positive. If the initial distribution is (v_1, v_2), the distribution after n steps alternates between (v_1, v_2) and (v_2, v_1). ◀

Another example that fails to satisfy the conditions of Theorem 3.10.4 is the gambler's ruin problem from Sec. 2.4.

Example 3.10.16

Gambler's Ruin. In Sec. 2.4, we described the gambler's ruin problem, in which a gambler wins one dollar with probability p and loses one dollar with probability $1 - p$ on each play of a game. The sequence of amounts held by the gambler through the course of those plays forms a Markov chain with two absorbing states, namely, 0 and k. There are $k - 1$ other states, namely, $1, \ldots, k - 1$. (This notation violates our use of k to stand for the number of states, which is $k + 1$ in this example. We felt this was less confusing than switching from the original notation of Sec. 2.4.) The transition matrix has first and last row being $(1, 0, \ldots, 0)$ and $(0, \ldots, 1)$, respectively. The ith row (for $i = 1, \ldots, k - 1$) has 0 everywhere except in coordinate $i - 1$ where it has $1 - p$ and in coordinate $i + 1$ where it has p. Unlike Example 3.10.15, this time the sequence of matrices $\boldsymbol{P}^m$ converges but there is no unique stationary distribution. The limit of $\boldsymbol{P}^m$ has as its last column the numbers $a_0, \ldots, a_k$, where a_i is the probability that the fortune of a gambler who starts with i dollars reaches k dollars before it reaches 0 dollars. The first column of the limit has the numbers $1 - a_0, \ldots, 1 - a_k$ and the rest of the limit matrix is all zeros. The stationary distributions have the same form as those in Example 3.10.13, namely, all probability is in the absorbing states. ◀

Summary

A Markov chain is a stochastic process, a sequence of random variables giving the states of the process, in which the conditional distribution of the state at the next time given all of the past states depends on the past states only through the most recent state. For Markov chains with finitely many states and stationary transition distributions, the transitions over time can be described by a matrix giving the probabilities of transition from the state indexing the row to the state indexing the column (the transition matrix $\boldsymbol{P}$). The initial probability vector $\boldsymbol{v}$ gives the distribution of the state at time 1. The transition matrix and initial probability vector together allow calculation of all probabilities associated with the Markov chain. In particular, $\boldsymbol{P}^n$ gives the probabilities of transitions over n time periods, and $\boldsymbol{v}\boldsymbol{P}^n$ gives the distribution of the state at time $n + 1$. A stationary distribution is a probability vector $\boldsymbol{v}$ such that $\boldsymbol{v}\boldsymbol{P} = \boldsymbol{v}$. Every finite Markov chain with stationary transition distributions has at least one stationary distribution. For many Markov chains, there is a unique stationary distribution and the distribution of the chain after n transitions converges to the stationary distribution as n goes to ∞.

Exercises

1. Consider the Markov chain in Example 3.10.2 with initial probability vector $\boldsymbol{v} = (1/2, 1/2)$.

a. Find the probability vector specifying the probabilities of the states at time $n = 2$.

b. Find the two-step transition matrix.

2. Suppose that the weather can be only sunny or cloudy and the weather conditions on successive mornings form a Markov chain with stationary transition probabilities. Suppose also that the transition matrix is as follows:

	Sunny	Cloudy
Sunny	0.7	0.3
Cloudy	0.6	0.4

a. If it is cloudy on a given day, what is the probability that it will also be cloudy the next day?

b. If it is sunny on a given day, what is the probability that it will be sunny on the next two days?

c. If it is cloudy on a given day, what is the probability that it will be sunny on at least one of the next three days?

3. Consider again the Markov chain described in Exercise 2.

a. If it is sunny on a certain Wednesday, what is the probability that it will be sunny on the following Saturday?

b. If it is cloudy on a certain Wednesday, what is the probability that it will be sunny on the following Saturday?

4. Consider again the conditions of Exercises 2 and 3.

a. If it is sunny on a certain Wednesday, what is the probability that it will be sunny on both the following Saturday and Sunday?

b. If it is cloudy on a certain Wednesday, what is the probability that it will be sunny on both the following Saturday and Sunday?

5. Consider again the Markov chain described in Exercise 2. Suppose that the probability that it will be sunny on a certain Wednesday is 0.2 and the probability that it will be cloudy is 0.8.

a. Determine the probability that it will be cloudy on the next day, Thursday.

b. Determine the probability that it will be cloudy on Friday.

c. Determine the probability that it will be cloudy on Saturday.

6. Suppose that a student will be either on time or late for a particular class and that the events that he is on time or late for the class on successive days form a Markov chain with stationary transition probabilities. Suppose also that if he is late on a given day, then the probability that he will be on time the next day is 0.8. Furthermore, if he is on time on a given day, then the probability that he will be late the next day is 0.5.

a. If the student is late on a certain day, what is the probability that he will be on time on each of the next three days?

b. If the student is on time on a given day, what is the probability that he will be late on each of the next three days?

7. Consider again the Markov chain described in Exercise 6.

a. If the student is late on the first day of class, what is the probability that he will be on time on the fourth day of class?

b. If the student is on time on the first day of class, what is the probability that he will be on time on the fourth day of class?

8. Consider again the conditions of Exercises 6 and 7. Suppose that the probability that the student will be late on the first day of class is 0.7 and that the probability that he will be on time is 0.3.

a. Determine the probability that he will be late on the second day of class.

b. Determine the probability that he will be on time on the fourth day of class.

9. Suppose that a Markov chain has four states 1, 2, 3, 4 and stationary transition probabilities as specified by the following transition matrix:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1/4 & 1/4 & 0 & 1/2 \\ 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \end{array}.$$

a. If the chain is in state 3 at a given time n, what is the probability that it will be in state 2 at time $n + 2$?

b. If the chain is in state 1 at a given time n, what is the probability that it will be in state 3 at time $n + 3$?

10. Let X_1 denote the initial state at time 1 of the Markov chain for which the transition matrix is as specified in Exercise 5, and suppose that the initial probabilities are as follows:

$$\Pr(X_1 = 1) = 1/8, \Pr(X_1 = 2) = 1/4,$$
$$\Pr(X_1 = 3) = 3/8, \Pr(X_1 = 4) = 1/4.$$

Determine the probabilities that the chain will be in states 1, 2, 3, and 4 at time n for each of the following values of n: **(a)** $n = 2$; **(b)** $n = 3$; **(c)** $n = 4$.

11. Each time that a shopper purchases a tube of toothpaste, she chooses either brand A or brand B. Suppose that the probability is 1/3 that she will choose the same brand

chosen on her previous purchase, and the probability is 2/3 that she will switch brands.

a. If her first purchase is brand A, what is the probability that her fifth purchase will be brand B?

b. If her first purchase is brand B, what is the probability that her fifth purchase will be brand B?

12. Suppose that three boys A, B, and C are throwing a ball from one to another. Whenever A has the ball, he throws it to B with a probability of 0.2 and to C with a probability of 0.8. Whenever B has the ball, he throws it to A with a probability of 0.6 and to C with a probability of 0.4. Whenever C has the ball, he is equally likely to throw it to either A or B.

a. Consider this process to be a Markov chain and construct the transition matrix.

b. If each of the three boys is equally likely to have the ball at a certain time n, which boy is most likely to have the ball at time $n + 2$?

13. Suppose that a coin is tossed repeatedly in such a way that heads and tails are equally likely to appear on any given toss and that all tosses are independent, with the following exception: Whenever either three heads or three tails have been obtained on three successive tosses, then the outcome of the next toss is always of the opposite type. At time n ($n \geq 3$), let the state of this process be specified by the outcomes on tosses $n - 2$, $n - 1$, and n. Show that this process is a Markov chain with stationary transition probabilities and construct the transition matrix.

14. There are two boxes A and B, each containing red and green balls. Suppose that box A contains one red ball and two green balls and box B contains eight red balls and two green balls. Consider the following process: One ball is selected at random from box A, and one ball is selected at random from box B. The ball selected from box A is then placed in box B and the ball selected from box B is placed in box A. These operations are then repeated indefinitely. Show that the numbers of red balls in box A form a Markov chain with stationary transition probabilities, and construct the transition matrix of the Markov chain.

15. Verify the rows of the transition matrix in Example 3.10.6 that correspond to current states $\{AA, Aa\}$ and $\{Aa, aa\}$.

16. Let the initial probability vector in Example 3.10.6 be $\boldsymbol{v} = (1/16, 1/4, 1/8, 1/4, 1/4, 1/16)$. Find the probabilities of the six states after one generation.

17. Return to Example 3.10.6. Assume that the state at time $n - 1$ is $\{Aa, aa\}$.

a. Suppose that we learn that X_{n+1} is $\{AA, aa\}$. Find the conditional distribution of X_n. (That is, find all the probabilities for the possible states at time n given that the state at time $n + 1$ is $\{AA, aa\}$.)

b. Suppose that we learn that X_{n+1} is $\{aa, aa\}$. Find the conditional distribution of X_n.

18. Return to Example 3.10.13. Prove that the stationary distributions described there are the only stationary distributions for that Markov chain.

19. Find the unique stationary distribution for the Markov chain in Exercise 2.

20. The unique stationary distribution in Exercise 9 is $\mathbf{v} = (0, 1, 0, 0)$. This is an instance of the following general result: Suppose that a Markov chain has exactly one absorbing state. Suppose further that, for each non-absorbing state k, there is n such that the probability is positive of moving from state k to the absorbing state in n steps. Then the unique stationary distribution has probability 1 in the absorbing state. Prove this result.

3.11 Supplementary Exercises

1. Suppose that X and Y are independent random variables, that X has the uniform distribution on the integers 1, 2, 3, 4, 5 (discrete), and that Y has the uniform distribution on the interval $[0, 5]$ (continuous). Let Z be a random variable such that $Z = X$ with probability 1/2 and $Z = Y$ with probability 1/2. Sketch the c.d.f. of Z.

2. Suppose that X and Y are independent random variables. Suppose that X has a discrete distribution concentrated on finitely many distinct values with p.f. f_1. Suppose that Y has a continuous distribution with p.d.f. f_2. Let $Z = X + Y$. Show that Z has a continuous distribution and find its p.d.f. *Hint:* First find the conditional p.d.f. of Z given $X = x$.

3. Suppose that the random variable X has the following c.d.f.:

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{2}{5}x & \text{for } 0 < x \leq 1, \\ \frac{3}{5}x - \frac{1}{5} & \text{for } 1 < x \leq 2, \\ 1 & \text{for } x > 2. \end{cases}$$

Verify that X has a continuous distribution, and determine the p.d.f. of X.

4. Suppose that the random variable X has a continuous distribution with the following p.d.f.:

$$f(x) = \frac{1}{2}e^{-|x|} \quad \text{for } -\infty < x < \infty.$$

Determine the value x_0 such that $F(x_0) = 0.9$, where $F(x)$ is the c.d.f. of X.

5. Suppose that X_1 and X_2 are i.i.d. random variables, and that each has the uniform distribution on the interval [0, 1]. Evaluate $\Pr(X_1^2 + X_2^2 \le 1)$.

6. For each value of $p > 1$, let

$$c(p) = \sum_{x=1}^{\infty} \frac{1}{x^p}.$$

Suppose that the random variable X has a discrete distribution with the following p.f.:

$$f(x) = \frac{1}{c(p)x^p} \quad \text{for } x = 1, 2, \ldots.$$

a. For each fixed positive integer n, determine the probability that X will be divisible by n.

b. Determine the probability that X will be odd.

7. Suppose that X_1 and X_2 are i.i.d. random variables, each of which has the p.f. $f(x)$ specified in Exercise 6. Determine the probability that $X_1 + X_2$ will be even.

8. Suppose that an electronic system comprises four components, and let X_j denote the time until component j fails to operate ($j = 1, 2, 3, 4$). Suppose that X_1, X_2, X_3, and X_4 are i.i.d. random variables, each of which has a continuous distribution with c.d.f. $F(x)$. Suppose that the system will operate as long as both component 1 and at least one of the other three components operate. Determine the c.d.f. of the time until the system fails to operate.

9. Suppose that a box contains a large number of tacks and that the probability X that a particular tack will land with its point up when it is tossed varies from tack to tack in accordance with the following p.d.f.:

$$f(x) = \begin{cases} 2(1-x) & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a tack is selected at random from the box and that this tack is then tossed three times independently. Determine the probability that the tack will land with its point up on all three tosses.

10. Suppose that the radius X of a circle is a random variable having the following p.d.f.:

$$f(x) = \begin{cases} \frac{1}{8}(3x+1) & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the p.d.f. of the area of the circle.

11. Suppose that the random variable X has the following p.d.f.:

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

Construct a random variable $Y = r(X)$ that has the uniform distribution on the interval [0, 5].

12. Suppose that the 12 random variables $X_1, \ldots, X_{12}$ are i.i.d. and each has the uniform distribution on the interval [0, 20]. For $j = 0, 1, \ldots, 19$, let I_j denote the interval $(j, j+1)$. Determine the probability that none of the 20 disjoint intervals I_j will contain more than one of the random variables $X_1, \ldots, X_{12}$.

13. Suppose that the joint distribution of X and Y is uniform over a set A in the xy-plane. For which of the following sets A are X and Y independent?

a. A circle with a radius of 1 and with its center at the origin

b. A circle with a radius of 1 and with its center at the point (3, 5)

c. A square with vertices at the four points (1, 1), (1, −1), (−1, −1), and (−1, 1)

d. A rectangle with vertices at the four points (0, 0), (0, 3), (1, 3), and (1, 0)

e. A square with vertices at the four points (0, 0), (1, 1), (0, 2), and (−1, 1)

14. Suppose that X and Y are independent random variables with the following p.d.f.'s:

$$f_1(x) = \begin{cases} 1 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(y) = \begin{cases} 8y & \text{for } 0 < y < \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of $\Pr(X > Y)$.

15. Suppose that, on a particular day, two persons A and B arrive at a certain store independently of each other. Suppose that A remains in the store for 15 minutes and B remains in the store for 10 minutes. If the time of arrival of each person has the uniform distribution over the hour between 9:00 A.M. and 10:00 A.M., what is the probability that A and B will be in the store at the same time?

16. Suppose that X and Y have the following joint p.d.f.:

$$f(x, y) = \begin{cases} 2(x+y) & \text{for } 0 < x < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** $\Pr(X < 1/2)$; **(b)** the marginal p.d.f. of X; and **(c)** the conditional p.d.f. of Y given that $X = x$.

17. Suppose that X and Y are random variables. The marginal p.d.f. of X is

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, the conditional p.d.f. of Y given that $X = x$ is

$$g(y|x) = \begin{cases} \frac{3y^2}{x^3} & \text{for } 0 < y < x, \\ 0 & \text{otherwise.} \end{cases}$$

Determine **(a)** the marginal p.d.f. of Y and **(b)** the conditional p.d.f. of X given that $Y = y$.

18. Suppose that the joint distribution of X and Y is uniform over the region in the xy-plane bounded by the four lines $x = -1$, $x = 1$, $y = x + 1$, and $y = x - 1$. Determine **(a)** $\Pr(XY > 0)$ and **(b)** the conditional p.d.f. of Y given that $X = x$.

19. Suppose that the random variables X, Y, and Z have the following joint p.d.f.:

$$f(x, y, z) = \begin{cases} 6 & \text{for } 0 < x < y < z < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the univariate marginal p.d.f.'s of X, Y, and Z.

20. Suppose that the random variables X, Y, and Z have the following joint p.d.f.:

$$f(x, y, z) = \begin{cases} 2 & \text{for } 0 < x < y < 1 \text{ and } 0 < z < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Evaluate $\Pr(3X > Y | 1 < 4Z < 2)$.

21. Suppose that X and Y are i.i.d. random variables, and that each has the following p.d.f.:

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

Also, let $U = X/(X + Y)$ and $V = X + Y$.

a. Determine the joint p.d.f. of U and V.

b. Are U and V independent?

22. Suppose that the random variables X and Y have the following joint p.d.f.:

$$f(x, y) = \begin{cases} 8xy & \text{for } 0 \le x \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Also, let $U = X/Y$ and $V = Y$.

a. Determine the joint p.d.f. of U and V.

b. Are X and Y independent?

c. Are U and V independent?

23. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables, each having the following c.d.f.:

$$F(x) = \begin{cases} 0 & \text{for } x \le 0, \\ 1 - e^{-x} & \text{for } x > 0. \end{cases}$$

Let $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$. Determine the conditional p.d.f. of Y_1 given that $Y_n = y_n$.

24. Suppose that X_1, X_2, and X_3 form a random sample of three observations from a distribution having the following p.d.f.:

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the p.d.f. of the range of the sample.

25. In this exercise, we shall provide an approximate justification for Eq. (3.6.6). First, remember that if a and b are close together, then

$$\int_a^b r(t)dt \approx (b - a)r\left(\frac{a + b}{2}\right). \tag{3.11.1}$$

Throughout this problem, assume that X and Y have joint p.d.f. f.

a. Use (3.11.1) to approximate $\Pr(y - \epsilon < Y \le y + \epsilon)$.

b. Use (3.11.1) with $r(t) = f(s, t)$ for fixed s to approximate

$$\Pr(X \le x \text{ and } y - \epsilon < Y \le y + \epsilon) = \int_{-\infty}^{x} \int_{y-\epsilon}^{y+\epsilon} f(s, t)\, dt\, ds.$$

c. Show that the ratio of the approximation in part (b) to the approximation in part (a) is $\int_{-\infty}^{x} g_1(s|y)\, ds$.

26. Let X_1, X_2 be two independent random variables each with p.d.f. $f_1(x) = e^{-x}$ for $x > 0$ and $f_1(x) = 0$ for $x \le 0$. Let $Z = X_1 - X_2$ and $W = X_1/X_2$.

a. Find the joint p.d.f. of X_1 and Z.

b. Prove that the conditional p.d.f. of X_1 given $Z = 0$ is

$$g_1(x_1|0) = \begin{cases} 2e^{-2x_1} & \text{for } x_1 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

c. Find the joint p.d.f. of X_1 and W.

d. Prove that the conditional p.d.f. of X_1 given $W = 1$ is

$$h_1(x_1|1) = \begin{cases} 4x_1e^{-2x_1} & \text{for } x_1 > 0, \\ 0 & \text{otherwise.} \end{cases}$$

e. Notice that $\{Z = 0\} = \{W = 1\}$, but the conditional distribution of X_1 given $Z = 0$ is not the same as the conditional distribution of X_1 given $W = 1$. This discrepancy is known as the *Borel paradox*. In light of the discussion that begins on page 146 about how conditional p.d.f.'s are not like conditioning on events of probability 0, show how "Z very close to 0" is not the same as "W very close to 1." *Hint:* Draw a set of axes for x_1 and x_2, and draw the two sets $\{(x_1, x_2) : |x_1 - x_2| < \epsilon\}$ and $\{(x_1, x_2) : |x_1/x_2 - 1| < \epsilon\}$ and see how much different they are.

27. Three boys A, B, and C are playing table tennis. In each game, two of the boys play against each other and the third boy does not play. The winner of any given game n plays again in game $n+1$ against the boy who did not play in game n, and the loser of game n does not play in game $n+1$. The probability that A will beat B in any game that they play against each other is 0.3, the probability that A will beat C is 0.6, and the probability that B will beat C is 0.8. Represent this process as a Markov chain with stationary transition probabilities by defining the possible states and constructing the transition matrix.

28. Consider again the Markov chain described in Exercise 27. **(a)** Determine the probability that the two boys who play against each other in the first game will play against each other again in the fourth game. **(b)** Show that this probability does not depend on which two boys play in the first game.

29. Find the unique stationary distribution for the Markov chain in Exercise 27.

Probability and Statistics

Chapter

4 EXPECTATION

4.1 The Expectation of a Random Variable

The distribution of a random variable X contains all of the probabilistic information about X. The entire distribution of X, however, is usually too cumbersome for presenting this information. Summaries of the distribution, such as the average value, or expected value, can be useful for giving people an idea of where we expect X to be without trying to describe the entire distribution. The expected value also plays an important role in the approximation methods that arise in Chapter 6.

Expectation for a Discrete Distribution

Example 4.1.1

Fair Price for a Stock. An investor is considering whether or not to invest \$18 per share in a stock for one year. The value of the stock after one year, in dollars, will be $18 + X$, where X is the amount by which the price changes over the year. At present X is unknown, and the investor would like to compute an "average value" for X in order to compare the return she expects from the investment to what she would get by putting the \$18 in the bank at 5% interest. ◀

The idea of finding an average value as in Example 4.1.1 arises in many applications that involve a random variable. One popular choice is what we call the *mean* or *expected value* or *expectation*.

The intuitive idea of the mean of a random variable is that it is the weighted average of the possible values of the random variable with the weights equal to the probabilities.

Example 4.1.2

Stock Price Change. Suppose that the change in price of the stock in Example 4.1.1 is a random variable X that can assume only the four different values -2, 0, 1, and 4, and that $\Pr(X = -2) = 0.1$, $\Pr(X = 0) = 0.4$, $\Pr(X = 1) = 0.3$, and $\Pr(X = 4) = 0.2$. Then the weighted avarage of these values is

$$-2(0.1) + 0(0.4) + 1(0.3) + 4(0.2) = 0.9.$$

The investor now compares this with the interest that would be earned on \$18 at 5% for one year, which is $18 \times 0.05 = 0.9$ dollars. From this point of view, the price of \$18 seems fair. ◀

The calculation in Example 4.1.2 generalizes easily to every random variable that assumes only finitely many values. Possible problems arise with random variables that assume more than finitely many values, especially when the collection of possible values is unbounded.

Definition 4.1.1 Mean of Bounded Discrete Random Variable. Let X be a bounded discrete random variable whose p.f. is f. The *expectation of* X, denoted by $E(X)$, is a number defined as follows:

$$E(X) = \sum_{\text{All } x} x f(x). \tag{4.1.1}$$

The expectation of X is also referred to as the *mean of* X or the *expected value of* X.

In Example 4.1.2, $E(X) = 0.9$. Notice that 0.9 is not one of the possible values of X in that example. This is typically the case with discrete random variables.

Example 4.1.3 Bernoulli Random Variable. Let X have the Bernoulli distribution with parameter p, that is, assume that X takes only the two values 0 and 1 with $\Pr(X = 1) = p$. Then the mean of X is

$$E(X) = 0 \times (1 - p) + 1 \times p = p. \quad \blacktriangleleft$$

If X is unbounded, it might still be possible to define $E(X)$ as the weighted average of its possible values. However, some care is needed.

Definition 4.1.2 Mean of General Discrete Random Variable. Let X be a discrete random variable whose p.f. is f. Suppose that at least one of the following sums is finite:

$$\sum_{\text{Positive } x} x f(x), \qquad \sum_{\text{Negative } x} x f(x). \tag{4.1.2}$$

Then the *mean*, *expectation*, or *expected value* of X is said to *exist* and is defined to be

$$E(X) = \sum_{\text{All } x} x f(x). \tag{4.1.3}$$

If both of the sums in (4.1.2) are infinite, then $E(X)$ *does not exist*.

The reason that the expectation fails to exist if both of the sums in (4.1.2) are infinite is that, in such cases, the sum in (4.1.3) is not well-defined. It is known from calculus that the sum of an infinite series whose positive and negative terms both add to infinity either fails to converge or can be made to converge to many different values by rearranging the terms in different orders. We don't want the meaning of expected value to depend on arbitrary choices about what order to add numbers. If only one of two sums in (4.1.3) is infiinte, then the expected value is also infinite with the same sign as that of the sum that is infinite. If both sums are finite, then the sum in (4.1.3) converges and doesn't depend on the order in which the terms are added.

Example 4.1.4 The Mean of X Does Not Exist. Let X be a random variable whose p.f. is

$$f(x) = \begin{cases} \dfrac{1}{2|x|(|x|+1)} & \text{if } x = \pm 1, \pm 2, \pm 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

It can be verified that this function satisfies the conditions required to be a p.f. The two sums in (4.1.2) are

$$\sum_{x=-1}^{-\infty} x\frac{1}{2|x|(|x|+1)} = -\infty \quad \text{and} \quad \sum_{x=1}^{\infty} x\frac{1}{2x(x+1)} = \infty;$$

hence, $E(X)$ does not exist. ◀

Example 4.1.5 An Infinite Mean. Let X be a random variable whose p.f. is

$$f(x) = \begin{cases} \dfrac{1}{x(x+1)} & \text{if } x = 1, 2, 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The sum over negative values in Eq. (4.1.2) is 0, so the mean of X exists and is

$$E(X) = \sum_{x=1}^{\infty} x\frac{1}{x(x+1)} = \infty.$$

We say that the mean of X is *infinite* in this case. ◀

Note: The Expectation of X Depends Only on the Distribution of X. Although $E(X)$ is called the expectation of X, it depends only on the distribution of X. Every two random variables that have the same distribution will have the same expectation even if they have nothing to do with each other. For this reason, we shall often refer to the expectation of a distribution even if we do not have in mind a random variable with that distribution.

Expectation for a Continuous Distribution

The idea of computing a weighted average of the possible values can be generalized to continuous random variables by using integrals instead of sums. The distinction between bounded and unbounded random variables arises in this case for the same reasons.

Definition 4.1.3 Mean of Bounded Continuous Random Variable. Let X be a bounded continuous random variable whose p.d.f. is f. The *expectation of* X, denoted $E(X)$, is defined as follows:

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx. \tag{4.1.4}$$

Once again, the expectation is also called the *mean* or the *expected value*.

Example 4.1.6 Expected Failure Time. An appliance has a maximum lifetime of one year. The time X until it fails is a random variable with a continuous distribution having p.d.f.

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E(X) = \int_0^1 x(2x)\,dx = \int_0^1 2x^2\,dx = \frac{2}{3}.$$

We can also say that the expectation of the distribution with p.d.f. f is 2/3. ◀

For general continuous random variables, we modify Definition 4.1.2.

Definition 4.1.4 Mean of General Continuous Random Variable. Let X be a continuous random variable whose p.d.f. is f. Suppose that at least one of the following integrals is finite:

$$\int_0^\infty xf(x)dx, \quad \int_{-\infty}^0 xf(x)dx. \tag{4.1.5}$$

Then the *mean*, *expectation*, or *expected value* of X is said to *exist* and is defined to be

$$E(X) = \int_{-\infty}^\infty xf(x)dx. \tag{4.1.6}$$

If both of the integrals in (4.1.5) are infinite, then $E(X)$ *does not exist*.

Example 4.1.7 Failure after Warranty. A product has a warranty of one year. Let X be the time at which the product fails. Suppose that X has a continuous distribution with the p.d.f.

$$f(x) = \begin{cases} 0 & \text{for } x < 1, \\ \frac{2}{x^3} & \text{for } x \geq 1. \end{cases}$$

The expected time to failure is then

$$E(X) = \int_1^\infty x\frac{2}{x^3}dx = \int_1^\infty \frac{2}{x^2}dx = 2. \quad \blacktriangleleft$$

Example 4.1.8 A Mean That Does Not Exist. Suppose that a random variable X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } -\infty < x < \infty. \tag{4.1.7}$$

This distribution is called the *Cauchy distribution*. We can verify the fact that $\int_{-\infty}^\infty f(x)\,dx = 1$ by using the following standard result from elementary calculus:

$$\frac{d}{dx}\tan^{-1}x = \frac{1}{1+x^2} \quad \text{for } -\infty < x < \infty.$$

The two integrals in (4.1.5) are

$$\int_0^\infty \frac{x}{\pi(1+x^2)}dx = \infty \quad \text{and} \quad \int_{-\infty}^0 \frac{x}{\pi(1+x^2)}dx = -\infty;$$

hence, the mean of X does not exist. $\blacktriangleleft$

Interpretation of the Expectation

Relation of the Mean to the Center of Gravity The expectation of a random variable or, equivalently, the mean of its distribution can be regarded as being the center of gravity of that distribution. To illustrate this concept, consider, for example, the p.f. sketched in Fig. 4.1. The x-axis may be regarded as a long weightless rod to which weights are attached. If a weight equal to $f(x_j)$ is attached to this rod at each point x_j, then the rod will be balanced if it is supported at the point $E(X)$.

Now consider the p.d.f. sketched in Fig. 4.2. In this case, the x-axis may be regarded as a long rod over which the mass varies continuously. If the density of

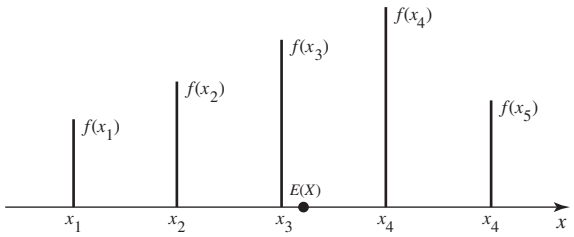

Figure 4.1 The mean of a discrete distribution.

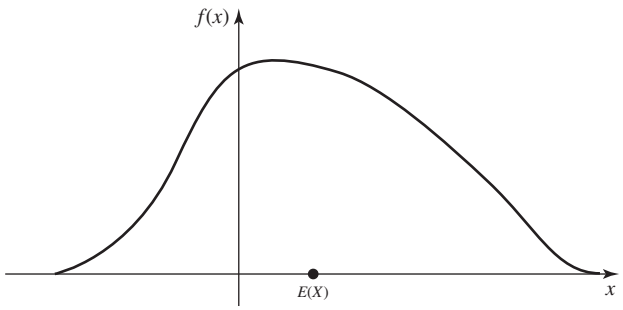

Figure 4.2 The mean of a continuous distribution.

the rod at each point x is equal to $f(x)$, then the center of gravity of the rod will be located at the point $E(X)$, and the rod will be balanced if it is supported at that point.

It can be seen from this discussion that the mean of a distribution can be affected greatly by even a very small change in the amount of probability that is assigned to a large value of x. For example, the mean of the distribution represented by the p.f. in Fig. 4.1 can be moved to any specified point on the x-axis, no matter how far from the origin that point may be, by removing an arbitrarily small but positive amount of probability from one of the points x_j and adding this amount of probability at a point far enough from the origin.

Suppose now that the p.f. or p.d.f. f of some distribution is symmetric with respect to a given point x_0 on the x-axis. In other words, suppose that $f(x_0 + \delta) = f(x_0 - \delta)$ for all values of δ. Also assume that the mean $E(X)$ of this distribution exists. In accordance with the interpretation that the mean is at the center of gravity, it follows that $E(X)$ must be equal to x_0, which is the point of symmetry. The following example emphasizes the fact that it is necessary to make certain that the mean $E(X)$ exists before it can be concluded that $E(X) = x_0$.

Example 4.1.9

The Cauchy Distribution. Consider again the p.d.f. specified by Eq. (4.1.7), which is sketched in Fig. 4.3. This p.d.f. is symmetric with respect to the point $x = 0$. Therefore, if the mean of the Cauchy distribution existed, its value would have to be 0. However, we saw in Example 4.1.8 that the mean of X does not exist.

The reason for the nonexistence of the mean of the Cauchy distribution is as follows: When the curve $y = f(x)$ is sketched as in Fig. 4.3, its tails approach the x-axis rapidly enough to permit the total area under the curve to be equal to 1. On the other hand, if each value of $f(x)$ is multiplied by x and the curve $y = xf(x)$ is sketched, as in Fig. 4.4, the tails of this curve approach the x-axis so slowly that the total area between the x-axis and each part of the curve is infinite. ◀

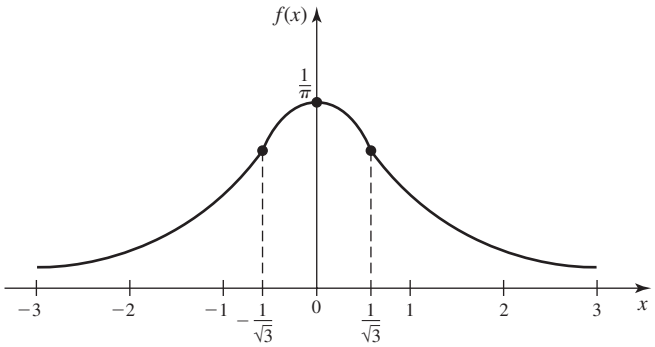

Figure 4.3 The p.d.f. of a Cauchy distribution.

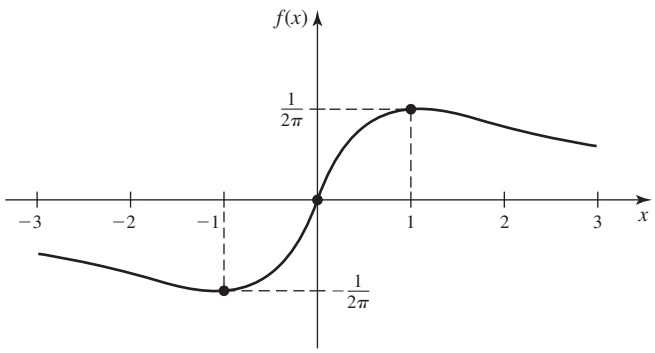

Figure 4.4 The curve $y = xf(x)$ for the Cauchy distribution.

The Expectation of a Function

Example 4.1.10

Failure Rate and Time to Failure. Suppose that appliances manufactured by a particular company fail at a rate of X per year, where X is currently unknown and hence is a random variable. If we are interested in predicting how long such an appliance will last before failure, we might use the mean of $1/X$. How can we calculate the mean of $Y = 1/X$? ◀

Functions of a Single Random Variable If X is a random variable for which the p.d.f. is f, then the expectation of each real-valued function $r(X)$ can be found by applying the definition of expectation to the distribution of $r(X)$ as follows: Let $Y = r(X)$, determine the probability distribution of Y, and then determine $E(Y)$ by applying either Eq. (4.1.1) or Eq. (4.1.4). For example, suppose that Y has a continuous distribution with the p.d.f. g. Then

$$E[r(X)] = E(Y) = \int_{-\infty}^{\infty} yg(y)\,dy, \tag{4.1.8}$$

if the expectation exists.

Example 4.1.11

Failure Rate and Time to Failure. In Example 4.1.10, suppose that the p.d.f. of X is

$$f(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $r(x) = 1/x$. Using the methods of Sec. 3.8, we can find the p.d.f. of $Y = r(X)$ as

$$g(y) = \begin{cases} 3y^{-4} & \text{if } y > 1, \\ 0 & \text{otherwise.} \end{cases}$$

The mean of Y is then

$$E(Y) = \int_0^\infty y 3y^{-4} dy = \frac{3}{2}. \quad \blacktriangleleft$$

Although the method of Example 4.1.11 can be used to find the mean of a continuous random variable, it is not actually necessary to determine the p.d.f. of $r(X)$ in order to calculate the expectation $E[r(X)]$. In fact, it can be shown that the value of $E[r(X)]$ can always be calculated directly using the following result.

Theorem 4.1.1 **Law of the Unconscious Statistician.** Let X be a random variable, and let r be a real-valued function of a real variable. If X has a continuous distribution, then

$$E[r(X)] = \int_{-\infty}^{\infty} r(x) f(x)\, dx, \tag{4.1.9}$$

if the mean exists. If X has a discrete distribution, then

$$E[r(X)] = \sum_{\text{All } x} r(x) f(x), \tag{4.1.10}$$

if the mean exists.

Proof A general proof will not be given here. However, we shall provide a proof for two special cases. First, suppose that the distribution of X is discrete. Then the distribution of Y must also be discrete. Let g be the p.f. of Y. For this case,

$$\begin{aligned} \sum_y y g(y) &= \sum_y y \Pr[r(X) = y] \\ &= \sum_y y \sum_{x:r(x)=y} f(x) \\ &= \sum_y \sum_{x:r(x)=y} r(x) f(x) = \sum_x r(x) f(x). \end{aligned}$$

Hence, Eq. (4.1.10) yields the same value as one would obtain from Definition 4.1.1 applied to Y.

Second, suppose that the distribution of X is continuous. Suppose also, as in Sec. 3.8, that $r(x)$ is either strictly increasing or strictly decreasing with differentiable inverse $s(y)$. Then, if we change variables in Eq. (4.1.9) from x to $y = r(x)$,

$$\int_{-\infty}^{\infty} r(x) f(x)\, dx = \int_{-\infty}^{\infty} y f[s(y)] \left| \frac{ds(y)}{dy} \right| dy.$$

It now follows from Eq. (3.8.3) that the right side of this equation is equal to

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

Hence, Eqs. (4.1.8) and (4.1.9) yield the same value. ■

Theorem 4.1.1 is called the law of the unconscious statistician because many people treat Eqs. (4.1.9) and (4.1.10) as the definition of $E[r(X)]$ and forget that the definition of the mean of $Y = r(X)$ is given in Definitions 4.1.2 and 4.1.4.

Example 4.1.12 Failure Rate and Time to Failure. In Example 4.1.11, we can apply Theorem 4.1.1 to find

$$E(Y) = \int_0^1 \frac{1}{x} 3x^2 dx = \frac{3}{2},$$

the same result we got in Example 4.1.11. ◀

Example 4.1.13 Determining the Expectation of $X^{1/2}$. Suppose that the p.d.f. of X is as given in Example 4.1.6 and that $Y = X^{1/2}$. Then, by Eq. (4.1.9),

$$E(Y) = \int_0^1 x^{1/2}(2x)\, dx = 2\int_0^1 x^{3/2}\, dx = \frac{4}{5}.$$ ◀

Note: In General, $E[g(X)] \neq g(E(X))$. In Example 4.1.13, the mean of $X^{1/2}$ is 4/5. The mean of X was computed in Example 4.1.6 as 2/3. Note that $4/5 \neq (2/3)^{1/2}$. In fact, unless g is a linear function, it is generally the case that $E[g(X)] \neq g(E(X))$. A linear function g does satisfy $E[g(X)] = g(E(X))$, as we shall see in Theorem 4.2.1.

Example 4.1.14 Option Pricing. Suppose that common stock in the up-and-coming company A is currently priced at \$200 per share. As an incentive to get you to work for company A, you might be offered an option to buy a certain number of shares of the stock, one year from now, at a price of \$200. This could be quite valuable if you believed that the stock was very likely to rise in price over the next year. For simplicity, suppose that the price X of the stock one year from now is a discrete random variable that can take only two values (in dollars): 260 and 180. Let p be the probability that $X = 260$. You want to calculate the value of these stock options, either because you contemplate the possibility of selling them or because you want to compare Company A's offer to what other companies are offering. Let Y be the value of the option for one share when it expires in one year. Since nobody would pay \$200 for the stock if the price X is less than \$200, the value of the stock option is 0 if $X = 180$. If $X = 260$, one could buy the stock for \$200 per share and then immediately sell it for \$260. This brings in a profit of \$60 per share. (For simplicity, we shall ignore dividends and the transaction costs of buying and selling stocks.) Then $Y = h(X)$ where

$$h(x) = \begin{cases} 0 & \text{if } x = 180, \\ 60 & \text{if } x = 260. \end{cases}$$

Assume that an investor could earn 4% risk-free on any money invested for this same year. (Assume that the 4% includes any compounding.) If no other investment options were available, a fair cost of the option would then be what is called the *present value* of $E(Y)$ in one year. This equals the value c such that $E(Y) = 1.04c$. That is, the expected value of the option equals the amount of money the investor would have after one year without buying the option. We can find $E(Y)$ easily:

$$E(Y) = 0 \times (1 - p) + 60 \times p = 60p.$$

So, the fair price of an option to buy one share would be $c = 60p/1.04 = 57.69p$.

How should one determine the probability p? There is a standard method used in the finance industry for choosing p in this example. That method is to assume that

the present value of the mean of X (the stock price in one year) is equal to the current value of the stock price. That is, assume that the expected value of buying one share of stock and waiting one year to sell is the same as the result of investing the current cost of the stock risk-free for one year (multiplying by 1.04 in this example). In our example, this means $E(X) = 200 \times 1.04$. Since $E(X) = 260p + 180(1 - p)$, we set

$$200 \times 1.04 = 260p + 180(1 - p),$$

and obtain $p = 0.35$. The resulting price of an option to buy one share for \$200 in one year would be $\$57.69 \times 0.35 = \$20.19$. This price is called the *risk-neutral price of the option*. One can prove (see Exercise 14 in this section) that any price other than \$20.19 for the option would lead to unpleasant consequences in the market. ◀

Functions of Several Random Variables

Example 4.1.15 The Expectation of a Function of Two Variables. Let X and Y have a joint p.d.f., and suppose that we want the mean of $X^2 + Y^2$. The most straightforward but most difficult way to do this would be to use the methods of Sec. 3.9 to find the distribution of $Z = X^2 + Y^2$ and then apply the definition of mean to Z. ◀

There is a version of Theorem 4.1.1 for functions of more than one random variable. Its proof is not given here.

Theorem 4.1.2 Law of the Unconscious Statistician. Suppose that $X_1, \ldots, X_n$ are random variables with the joint p.d.f. $f(x_1, \ldots, x_n)$. Let r be a real-valued function of n real variables, and suppose that $Y = r(X_1, \ldots, X_n)$. Then $E(Y)$ can be determined directly from the relation

$$E(Y) = \int \cdots \int_{R^n} r(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n,$$

if the mean exists. Similarly, if $X_1, \ldots, X_n$ have a discrete joint distribution with p.f. $f(x_1, \ldots, x_n)$, the mean of $Y = r(X_1, \ldots, X_n)$ is

$$E(Y) = \sum_{\text{All } x_1, \ldots, x_n} r(x_1, \ldots, x_n) f(x_1, \ldots, x_n),$$

if the mean exists. ■

Example 4.1.16 Determining the Expectation of a Function of Two Variables. Suppose that a point (X, Y) is chosen at random from the square S containing all points (x, y) such that $0 \le x \le 1$ and $0 \le y \le 1$. We shall determine the expected value of $X^2 + Y^2$.

Since X and Y have the uniform distribution over the square S, and since the area of S is 1, the joint p.d.f. of X and Y is

$$f(x, y) = \begin{cases} 1 & \text{for } (x, y) \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} E(X^2 + Y^2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x^2 + y^2) f(x, y)\, dx\, dy \\ &= \int_0^1 \int_0^1 (x^2 + y^2)\, dx\, dy = \frac{2}{3}. \end{aligned}$$ ◀

Note: More General Distributions. In Example 3.2.7, we introduced a type of distribution that was neither discrete nor continuous. It is possible to define expectations for such distributions also. The definition is rather cumbersome, and we shall not pursue it here.

Summary

The expectation, expected value, or mean of a random variable is a summary of its distribution. If the probability distribution is thought of as a distribution of mass along the real line, then the mean is the center of mass. The mean of a function r of a random variable X can be calculated directly from the distribution of X without first finding the distribution of $r(X)$. Similarly, the mean of a function of a random vector $\boldsymbol{X}$ can be calculated directly from the distribution of $\boldsymbol{X}$.

Exercises

1. Suppose that X has the uniform distribution on the interval $[a, b]$. Find the mean of X.

2. If an integer between 1 and 100 is to be chosen at random, what is the expected value?

3. In a class of 50 students, the number of students n_i of each age i is shown in the following table:

Age i	n_i
18	20
19	22
20	4
21	3
25	1

If a student is to be selected at random from the class, what is the expected value of his age?

4. Suppose that one word is to be selected at random from the sentence THE GIRL PUT ON HER BEAUTIFUL RED HAT. If X denotes the number of letters in the word that is selected, what is the value of $E(X)$?

5. Suppose that one letter is to be selected at random from the 30 letters in the sentence given in Exercise 4. If Y denotes the number of letters in the word in which the selected letter appears, what is the value of $E(Y)$?

6. Suppose that a random variable X has a continuous distribution with the p.d.f. f given in Example 4.1.6. Find the expectation of $1/X$.

7. Suppose that a random variable X has the uniform distribution on the interval [0, 1]. Show that the expectation of $1/X$ is infinite.

8. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x, y) = \begin{cases} 12y^2 & \text{for } 0 \le y \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of $E(XY)$.

9. Suppose that a point is chosen at random on a stick of unit length and that the stick is broken into two pieces at that point. Find the expected value of the length of the longer piece.

10. Suppose that a particle is released at the origin of the xy-plane and travels into the half-plane where $x > 0$. Suppose that the particle travels in a straight line and that the angle between the positive half of the x-axis and this line is α, which can be either positive or negative. Suppose, finally, that the angle α has the uniform distribution on the interval $[-\pi/2, \pi/2]$. Let Y be the ordinate of the point at which the particle hits the vertical line $x = 1$. Show that the distribution of Y is a Cauchy distribution.

11. Suppose that the random variables $X_1, \ldots, X_n$ form a random sample of size n from the uniform distribution on the interval [0, 1]. Let $Y_1 = \min\{X_1, \ldots, X_n\}$, and let $Y_n = \max\{X_1, \ldots, X_n\}$. Find $E(Y_1)$ and $E(Y_n)$.

12. Suppose that the random variables $X_1, \ldots, X_n$ form a random sample of size n from a continuous distribution for which the c.d.f. is F, and let the random variables Y_1 and Y_n be defined as in Exercise 11. Find $E[F(Y_1)]$ and $E[F(Y_n)]$.

13. A stock currently sells for \$110 per share. Let the price of the stock at the end of a one-year period be X, which will take one of the values \$100 or \$300. Suppose that you have the option to buy shares of this stock at \$150 per share at the end of that one-year period. Suppose that money

could earn 5.8% risk-free over that one-year period. Find the risk-neutral price for the option to buy one share.

14. Consider the situation of pricing a stock option as in Example 4.1.14. We want to prove that a price other than \$20.19 for the option to buy one share in one year for \$200 would be unfair in some way.

a. Suppose that an investor (who has several shares of the stock already) makes the following transactions. She buys three more shares of the stock at \$200 per share and sells four options for \$20.19 each. The investor must borrow the extra \$519.24 necessary to make these transactions at 4% for the year. At the end of the year, our investor might have to sell four shares for \$200 each to the person who bought the options. In any event, she sells enough stock to pay back the amount borrowed plus the 4 percent interest. Prove that the investor has the same net worth (within rounding error) at the end of the year as she would have had without making these transactions, no matter what happens to the stock price. (A combination of stocks and options that produces no change in net worth is called a *risk-free portfolio*.)

b. Consider the same transactions as in part (a), but this time suppose that the option price is \$$x$ where $x < 20.19$. Prove that our investor loses $|4.16x - 84|$ dollars of net worth no matter what happens to the stock price.

c. Consider the same transactions as in part (a), but this time suppose that the option price is \$$x$ where $x > 20.19$. Prove that our investor gains $4.16x - 84$ dollars of net worth no matter what happens to the stock price.

The situations in parts (b) and (c) are called *arbitrage opportunities*. Such opportunities rarely exist for any length of time in financial markets. Imagine what would happen if the three shares and four options were changed to three million shares and four million options.

15. In Example 4.1.14, we showed how to price an option to buy one share of a stock at a particular price at a particular time in the future. This type of option is called a *call option*. A *put option* is an option to sell a share of a stock at a particular price \$$y$ at a particular time in the future. (If you don't own any shares when you wish to exercise the option, you can always buy one at the market price and then sell it for \$$y$.) The same sort of reasoning as in Example 4.1.14 could be used to price a put option. Consider the same stock as in Example 4.1.14 whose price in one year is X with the same distribution as in the example and the same risk-free interest rate. Find the risk-neutral price for an option to sell one share of that stock in one year at a price of \$220.

16. Let Y be a discrete random variable whose p.f. is the function f in Example 4.1.4. Let $X = |Y|$. Prove that the distribution of X has the p.d.f. in Example 4.1.5.

4.2 Properties of Expectations

In this section, we present some results that simplify the calculation of expectations for some common functions of random variables.

Basic Theorems

Suppose that X is a random variable for which the expectation $E(X)$ exists. We shall present several results pertaining to the basic properties of expectations.

Theorem 4.2.1 Linear Function. If $Y = aX + b$, where a and b are finite constants, then

$$E(Y) = aE(X) + b.$$

Proof We first shall assume, for convenience, that X has a continuous distribution for which the p.d.f. is f. Then

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

A similar proof can be given for a discrete distribution. ■

Example 4.2.1 Calculating the Expectation of a Linear Function. Suppose that $E(X) = 5$. Then

$$E(3X - 5) = 3E(X) - 5 = 10$$

and

$$E(-3X + 15) = -3E(X) + 15 = 0.$$ ◀

The following result follows from Theorem 4.2.1 with $a = 0$.

Corollary 4.2.1 If $X = c$ with probability 1, then $E(X) = c$. ■

Example 4.2.2 Investment. An investor is trying to choose between two possible stocks to buy for a three-month investment. One stock costs \$50 per share and has a rate of return of R_1 dollars per share for the three-month period, where R_1 is a random variable. The second stock costs \$30 per share and has a rate of return of R_2 per share for the same three-month period. The investor has a total of \$6000 to invest. For this example, suppose that the investor will buy shares of only one stock. (In Example 4.2.3, we shall consider strategies in which the investor buys more than one stock.) Suppose that R_1 has the uniform distribution on the interval $[-10, 20]$ and that R_2 has the uniform distribution on the interval $[-4.5, 10]$. We shall first compute the expected dollar value of investing in each of the two stocks. For the first stock, the \$6000 will purchase 120 shares, so the return will be $120R_1$, whose mean is $120E(R_1) = 600$. (Solve Exercise 1 in Sec. 4.1 to see why $E(R_1) = 5$.) For the second stock, the \$6000 will purchase 200 shares, so the return will be $200R_2$, whose mean is $200E(R_2) = 550$. The first stock has a higher expected return.

In addition to calculating expected return, we should also ask which of the two investments is riskier. We shall now compute the value at risk (VaR) at probability level 0.97 for each investment. (See Example 3.3.7 on page 113.) VaR will be the negative of the $1 - 0.97 = 0.03$ quantile for the return on each investment. For the first stock, the return $120R_1$ has the uniform distribution on the interval $[-1200, 2400]$ (see Exercise 14 in Sec. 3.8) whose 0.03 quantile is (according to Example 3.3.8 on page 114) $0.03 \times 2400 + 0.97 \times (-1200) = -1092$. So VaR$= 1092$. For the second stock, the return $200R_2$ has the uniform distribution on the interval $[-900, 2000]$ whose 0.03 quantile is $0.03 \times 2000 + 0.97 \times (-900) = -813$. So VaR$= 813$. Even though the first stock has higher expected return, the second stock seems to be slightly less risky in terms of VaR. How should we balance risk and expected return to choose between the two purchases? One way to answer this question is illustrated in Example 4.8.10, after we learn about utility. ◀

Theorem 4.2.2 If there exists a constant such that $\Pr(X \geq a) = 1$, then $E(X) \geq a$. If there exists a constant b such that $\Pr(X \leq b) = 1$, then $E(X) \leq b$.

Proof We shall assume again, for convenience, that X has a continuous distribution for which the p.d.f. is f, and we shall suppose first that $\Pr(X \geq a) = 1$. Because X is bounded below, the second integral in (4.1.5) is finite. Then

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{a}^{\infty} x f(x)\, dx$$
$$\geq \int_{a}^{\infty} a f(x)\, dx = a \Pr(X \geq a) = a.$$

The proof of the other part of the theorem and the proof for a discrete distribution are similar. ■

It follows from Theorem 4.2.2 that if $\Pr(a \le X \le b) = 1$, then $a \le E(X) \le b$.

Theorem 4.2.3 Suppose that $E(X) = a$ and that either $\Pr(X \ge a) = 1$ or $\Pr(X \le a) = 1$. Then $\Pr(X = a) = 1$.

Proof We shall provide a proof for the case in which X has a discrete distribution and $\Pr(X \ge a) = 1$. The other cases are similar. Let $x_1, x_2, \ldots$ include every value $x > a$ such that $\Pr(X = x) > 0$, if any. Let $p_0 = \Pr(X = a)$. Then,

$$E(X) = p_0 a + \sum_{j=1}^{\infty} x_j \Pr(X = x_j). \tag{4.2.1}$$

Each x_j in the sum on the right side of Eq. (4.2.1) is greater than a. If we replace all of the x_j's by a, the sum can't get larger, and hence

$$E(X) \ge p_0 a + \sum_{j=1}^{\infty} a \Pr(X = x_j) = a. \tag{4.2.2}$$

Furthermore, the inequality in Eq. (4.2.2) will be strict if there is even one $x > a$ with $\Pr(X = x) > 0$. This contradicts $E(X) = a$. Hence, there can be no $x > a$ such that $\Pr(X = x) > 0$. ■

Theorem 4.2.4 If $X_1, \ldots, X_n$ are n random variables such that each expectation $E(X_i)$ is finite $(i = 1, \ldots, n)$, then

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

Proof We shall first assume that $n = 2$ and also, for convenience, that X_1 and X_2 have a continuous joint distribution for which the joint p.d.f. is f. Then

$$\begin{aligned}
E(X_1 + X_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 + x_2) f(x_1, x_2)\, dx_1\, dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1, x_2)\, dx_1\, dx_2 + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f(x_1, x_2)\, dx_1\, dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1, x_2)\, dx_2\, dx_1 + \int_{-\infty}^{\infty} x_2 f_2(x_2)\, dx_2 \\
&= \int_{-\infty}^{\infty} x_1 f_1(x_1)\, dx_1 + \int_{-\infty}^{\infty} x_2 f_2(x_2)\, dx_2 \\
&= E(X_1) + E(X_2),
\end{aligned}$$

where f_1 and f_2 are the marginal p.d.f.'s of X_1 and X_2. The proof for a discrete distribution is similar. Finally, the theorem can be established for each positive integer n by an induction argument. ■

It should be emphasized that, in accordance with Theorem 4.2.4, the expectation of the sum of several random variables always equals the sum of their individual expectations, regardless of what their joint distribution is. Even though the joint p.d.f. of X_1 and X_2 appeared in the proof of Theorem 4.2.4, only the marginal p.d.f.'s figured into the calculation of $E(X_1 + X_2)$.

The next result follows easily from Theorems 4.2.1 and 4.2.4.

Corollary 4.2.2 Assume that $E(X_i)$ is finite for $i = 1, \ldots, n$. For all constants $a_1, \ldots, a_n$ and b,

$$E(a_1 X_1 + \cdots + a_n X_n + b) = a_1 E(X_1) + \cdots + a_n E(X_n) + b.$$ ■

Example 4.2.3 Investment Portfolio. Suppose that the investor with \$6000 in Example 4.2.2 can buy shares of both of the two stocks. Suppose that the investor buys s_1 shares of the first stock at \$50 per share and s_2 shares of the second stock at \$30 per share. Such a combination of investments is called a *portfolio*. Ignoring possible problems with fractional shares, the values of s_1 and s_2 must satisfy

$$50s_1 + 30s_2 = 6000,$$

in order to invest the entire \$6000. The return on this portfolio will be $s_1R_1 + s_2R_2$. The mean return will be

$$s_1E(R_1) + s_2E(R_2) = 5s_1 + 2.75s_2.$$

For example, if $s_1 = 54$ and $s_2 = 110$, then the mean return is 572.5. ◀

Example 4.2.4 Sampling without Replacement. Suppose that a box contains red balls and blue balls and that the proportion of red balls in the box is p ($0 \le p \le 1$). Suppose that n balls are selected from the box at random *without replacement*, and let X denote the number of red balls that are selected. We shall determine the value of $E(X)$.

We shall begin by defining n random variables $X_1, \ldots, X_n$ as follows: For $i = 1, \ldots, n$, let $X_i = 1$ if the ith ball that is selected is red, and let $X_i = 0$ if the ith ball is blue. Since the n balls are selected without replacement, the random variables $X_1, \ldots, X_n$ are dependent. However, the marginal distribution of each X_i can be derived easily (see Exercise 10 of Sec. 1.7). We can imagine that all the balls are arranged in the box in some random order, and that the first n balls in this arrangement are selected. Because of randomness, the probability that the ith ball in the arrangement will be red is simply p. Hence, for $i = 1, \ldots, n$,

$$\Pr(X_i = 1) = p \quad \text{and} \quad \Pr(X_i = 0) = 1 - p. \tag{4.2.3}$$

Therefore, $E(X_i) = 1(p) + 0(1 - p) = p$.

From the definition of $X_1, \ldots, X_n$, it follows that $X_1 + \cdots + X_n$ is equal to the total number of red balls that are selected. Therefore, $X = X_1 + \cdots + X_n$ and, by Theorem 4.2.4,

$$E(X) = E(X_1) + \cdots + E(X_n) = np. \tag{4.2.4}$$

◀

Note: In General, $E[g(X)] \neq g(E(X))$. Theorems 4.2.1 and 4.2.4 imply that if g is a linear function of a random vector $\boldsymbol{X}$, then $E[g(\boldsymbol{X})] = g(E(\boldsymbol{X}))$. For a nonlinear function g, we have already seen Example 4.1.13 in which $E[g(\boldsymbol{X})] \neq g(E(\boldsymbol{X}))$. Jensen's inequality (Theorem 4.2.5) gives a relationship between $E[g(\boldsymbol{X})]$ and $g(E(\boldsymbol{X}))$ for another special class of functions.

Definition 4.2.1 Convex Functions. A function g of a vector argument is *convex* if, for every $\alpha \in (0, 1)$, and every $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$g[\alpha\boldsymbol{x} + (1 - \alpha)\boldsymbol{y}] \ge \alpha g(\boldsymbol{x}) + (1 - \alpha)g(\boldsymbol{y}).$$

The proof of Theorem 4.2.5 is not given, but one special case is left to the reader in Exercise 13.

Theorem 4.2.5 Jensen's Inequality. Let g be a convex function, and let $\boldsymbol{X}$ be a random vector with finite mean. Then $E[g(\boldsymbol{X})] \ge g(E(\boldsymbol{X}))$. ■

Example 4.2.5

Sampling with Replacement. Suppose again that in a box containing red balls and blue balls, the proportion of red balls is p $(0 \le p \le 1)$. Suppose now, however, that a random sample of n balls is selected from the box *with replacement*. If X denotes the number of red balls in the sample, then X has the binomial distribution with parameters n and p, as described in Sec. 3.1. We shall now determine the value of $E(X)$.

As before, for $i = 1, \ldots, n$, let $X_i = 1$ if the ith ball that is selected is red, and let $X_i = 0$ otherwise. Then, as before, $X = X_1 + \cdots + X_n$. In this problem, the random variables $X_1, \ldots, X_n$ are independent, and the marginal distribution of each X_i is again given by Eq. (4.2.3). Therefore, $E(X_i) = p$ for $i = 1, \ldots, n$, and it follows from Theorem 4.2.4 that

$$E(X) = np. \tag{4.2.5}$$

Thus, the mean of the binomial distribution with parameters n and p is np. The p.f. $f(x)$ of this binomial distribution is given by Eq. (3.1.4), and the mean can be computed directly from the p.f. as follows:

$$E(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x}. \tag{4.2.6}$$

Hence, by Eq. (4.2.5), the value of the sum in Eq. (4.2.6) must be np. ◀

It is seen from Eqs. (4.2.4) and (4.2.5) that the expected number of red balls in a sample of n balls is np, regardless of whether the sample is selected with or without replacement. However, the distribution of the number of red balls is different depending on whether sampling is done with or without replacement (for $n > 1$). For example, $\Pr(X = n)$ is always smaller in Example 4.2.4 where sampling is done without replacement than in Example 4.2.5 where sampling is done with replacement, if $n > 1$. (See Exercise 27 in Sec. 4.9.)

Example 4.2.6

Expected Number of Matches. Suppose that a person types n letters, types the addresses on n envelopes, and then places each letter in an envelope in a random manner. Let X be the number of letters that are placed in the correct envelopes. We shall find the mean of X. (In Sec. 1.10, we did a more difficult calculation with this same example.)

For $i = 1, \ldots, n$, let $X_i = 1$ if the ith letter is placed in the correct envelope, and let $X_i = 0$ otherwise. Then, for $i = 1, \ldots, n$,

$$\Pr(X_i = 1) = \frac{1}{n} \quad \text{and} \quad \Pr(X_i = 0) = 1 - \frac{1}{n}.$$

Therefore,

$$E(X_i) = \frac{1}{n} \quad \text{for } i = 1, \ldots, n.$$

Since $X = X_1 + \cdots + X_n$, it follows that

$$\begin{aligned} E(X) &= E(X_1) + \cdots + E(X_n) \\ &= \frac{1}{n} + \cdots + \frac{1}{n} = 1. \end{aligned}$$

Thus, the expected value of the number of correct matches of letters and envelopes is 1, regardless of the value of n. ◀

Expectation of a Product of Independent Random Variables

Theorem 4.2.6 If $X_1, \ldots, X_n$ are n independent random variables such that each expectation $E(X_i)$ is finite ($i = 1, \ldots, n$), then

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i).$$

Proof We shall again assume, for convenience, that $X_1, \ldots, X_n$ have a continuous joint distribution for which the joint p.d.f. is f. Also, we shall let f_i denote the marginal p.d.f. of X_i ($i = 1, \ldots, n$). Then, since the variables $X_1, \ldots, X_n$ are independent, it follows that at every point $(x_1, \ldots, x_n) \in R^n$,

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i).$$

Therefore,

$$\begin{aligned} E\left(\prod_{i=1}^{n} X_i\right) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} x_i\right) f(x_1, \cdots, x_n)\, dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n} x_i f_i(x_i)\right] dx_1 \ldots dx_n \\ &= \prod_{i=1}^{n} \int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i = \prod_{i=1}^{n} E(X_i). \end{aligned}$$

The proof for a discrete distribution is similar. ■

The difference between Theorem 4.2.4 and Theorem 4.2.6 should be emphasized. If it is assumed that each expectation is finite, the expectation of the sum of a group of random variables is *always* equal to the sum of their individual expectations. However, the expectation of the product of a group of random variables is *not* always equal to the product of their individual expectations. If the random variables are *independent*, then this equality will also hold.

Example 4.2.7 Calculating the Expectation of a Combination of Random Variables. Suppose that X_1, X_2, and X_3 are independent random variables such that $E(X_i) = 0$ and $E(X_i^2) = 1$ for $i = 1, 2, 3$. We shall determine the value of $E[X_1^2(X_2 - 4X_3)^2]$.

Since X_1, X_2, and X_3 are independent, it follows that the two random variables X_1^2 and $(X_2 - 4X_3)^2$ are also independent. Therefore,

$$\begin{aligned} E[X_1^2(X_2 - 4X_3)^2] &= E(X_1^2)E[(X_2 - 4X_3)^2] \\ &= E(X_2^2 - 8X_2X_3 + 16X_3^2) \\ &= E(X_2^2) - 8E(X_2X_3) + 16E(X_3^2) \\ &= 1 - 8E(X_2)E(X_3) + 16 \\ &= 17. \end{aligned}$$

◀

Example 4.2.8 Repeated Filtering. A filtration process removes a random proportion of particulates in water to which it is applied. Suppose that a sample of water is subjected to this process twice. Let X_1 be the proportion of the particulates that are removed by the first pass. Let X_2 be the proportion of what remains after the first pass that

is removed by the second pass. Assume that X_1 and X_2 are independent random variables with common p.d.f. $f(x) = 4x^3$ for $0 < x < 1$ and $f(x) = 0$ otherwise. Let Y be the proportion of the original particulates that remain in the sample after two passes. Then $Y = (1 - X_1)(1 - X_2)$. Because X_1 and X_2 are independent, so too are $1 - X_1$ and $1 - X_2$. Since $1 - X_1$ and $1 - X_2$ have the same distribution, they have the same mean, call it μ. It follows that Y has mean μ^2. We can find μ as

$$\mu = E(1 - X_1) = \int_0^1 (1 - x_1)4x_1^3 dx_1 = 1 - \frac{4}{5} = 0.2.$$

It follows that $E(Y) = 0.2^2 = 0.04$. ◀

Expectation for Nonnegative Distributions

Theorem 4.2.7 Integer-Valued Random Variables. Let X be a random variable that can take only the values 0, 1, 2, ... Then

$$E(X) = \sum_{n=1}^{\infty} \Pr(X \geq n). \tag{4.2.7}$$

Proof First, we can write

$$E(X) = \sum_{n=0}^{\infty} n \Pr(X = n) = \sum_{n=1}^{\infty} n \Pr(X = n). \tag{4.2.8}$$

Next, consider the following triangular array of probabilities:

$$\begin{array}{llll} \Pr(X=1) & \Pr(X=2) & \Pr(X=3) & \cdots \\ & \Pr(X=2) & \Pr(X=3) & \cdots \\ & & \Pr(X=3) & \cdots \\ & & & \ddots \end{array}$$

We can compute the sum of all the elements in this array in two different ways because all of the summands are nonnegative. First, we can add the elements in each column of the array and then add these column totals. Thus, we obtain the value $\sum_{n=1}^{\infty} n \Pr(X = n)$. Second, we can add the elements in each row of the array and then add these row totals. In this way we obtain the value $\sum_{n=1}^{\infty} \Pr(X \geq n)$. Therefore,

$$\sum_{n=1}^{\infty} n \Pr(X = n) = \sum_{n=1}^{\infty} \Pr(X \geq n).$$

Eq. (4.2.7) now follows from Eq. (4.2.8). ■

Example 4.2.9 Expected Number of Trials. Suppose that a person repeatedly tries to perform a certain task until he is successful. Suppose also that the probability of success on each given trial is p $(0 < p < 1)$ and that all trials are independent. If X denotes the number of the trial on which the first success is obtained, then $E(X)$ can be determined as follows.

Since at least one trial is always required, $\Pr(X \geq 1) = 1$. Also, for $n = 2, 3, \ldots$, at least n trials will be required if and only if none of the first $n - 1$ trials results in success. Therefore,

$$\Pr(X \geq n) = (1 - p)^{n-1}.$$

By Eq. (4.2.7), it follows that

$$E(X) = 1 + (1-p) + (1-p)^2 + \cdots = \frac{1}{1-(1-p)} = \frac{1}{p}. \quad \blacktriangleleft$$

Theorem 4.2.7 has a more general version that applies to all nonnegative random variables.

Theorem 4.2.8 General Nonnegative Random Variable. Let X be a nonnegative random variable with c.d.f. F. Then

$$E(X) = \int_0^\infty [1 - F(x)]dx. \tag{4.2.9}$$

■

The proof of Theorem 4.2.8 is left to the reader in Exercises 1 and 2 in Sec. 4.9.

Example 4.2.10 Expected Waiting Time. Let X be the time that a customer spends waiting for service in a queue. Suppose that the c.d.f. of X is

$$F(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 - e^{-2x} & \text{if } x > 0. \end{cases}$$

Then the mean of X is

$$E(X) = \int_0^\infty e^{-2x}dx = \frac{1}{2}. \quad \blacktriangleleft$$

Summary

The mean of a linear function of a random vector is the linear function of the mean. In particular, the mean of a sum is the sum of the means. As an example, the mean of the binomial distribution with parameters n and p is np. No such relationship holds in general for nonlinear functions. For independent random variables, the mean of the product is the product of the means.

Exercises

1. Suppose that the return R (in dollars per share) of a stock has the uniform distribution on the interval $[-3, 7]$. Suppose also, that each share of the stock costs \$1.50. Let Y be the net return (total return minus cost) on an investment of 10 shares of the stock. Compute $E(Y)$.

2. Suppose that three random variables X_1, X_2, X_3 form a random sample from a distribution for which the mean is 5. Determine the value of

$$E(2X_1 - 3X_2 + X_3 - 4).$$

3. Suppose that three random variables X_1, X_2, X_3 form a random sample from the uniform distribution on the interval $[0, 1]$. Determine the value of

$$E[(X_1 - 2X_2 + X_3)^2].$$

4. Suppose that the random variable X has the uniform distribution on the interval $[0, 1]$, that the random variable Y has the uniform distribution on the interval $[5, 9]$, and that X and Y are independent. Suppose also that a rectangle is to be constructed for which the lengths of two adjacent sides are X and Y. Determine the expected value of the area of the rectangle.

5. Suppose that the variables $X_1, \ldots, X_n$ form a random sample of size n from a given continuous distribution on the real line for which the p.d.f. is f. Find the expectation of the number of observations in the sample that fall within a specified interval $a \le x \le b$.

6. Suppose that a particle starts at the origin of the real line and moves along the line in jumps of one unit. For each jump, the probability is p ($0 \le p \le 1$) that the particle will jump one unit to the left and the probability is $1 - p$ that the particle will jump one unit to the right. Find the expected value of the position of the particle after n jumps.

7. Suppose that on each play of a certain game a gambler is equally likely to win or to lose. Suppose that when he wins, his fortune is doubled, and that when he loses, his fortune is cut in half. If he begins playing with a given fortune c, what is the expected value of his fortune after n independent plays of the game?

8. Suppose that a class contains 10 boys and 15 girls, and suppose that eight students are to be selected at random from the class without replacement. Let X denote the number of boys that are selected, and let Y denote the number of girls that are selected. Find $E(X - Y)$.

9. Suppose that the proportion of defective items in a large lot is p, and suppose that a random sample of n items is selected from the lot. Let X denote the number of defective items in the sample, and let Y denote the number of nondefective items. Find $E(X - Y)$.

10. Suppose that a fair coin is tossed repeatedly until a head is obtained for the first time. **(a)** What is the expected number of tosses that will be required? **(b)** What is the expected number of tails that will be obtained before the first head is obtained?

11. Suppose that a fair coin is tossed repeatedly until exactly k heads have been obtained. Determine the expected number of tosses that will be required. *Hint:* Represent the total number of tosses X in the form $X = X_1 + \cdots + X_k$, where X_i is the number of tosses required to obtain the ith head after $i - 1$ heads have been obtained.

12. Suppose that the two return random variables R_1 and R_2 in Examples 4.2.2 and 4.2.3 are independent. Consider the portfolio at the end of Example 4.2.3 with $s_1 = 54$ shares of the first stock and $s_2 = 110$ shares of the second stock.

a. Prove that the change in value X of the portfolio has the p.d.f.

$$f(x) = \begin{cases} 3.87 \times 10^{-7}(x + 1035) & \text{if } -1035 < x < 560, \\ 6.1728 \times 10^{-4} & \text{if } 560 \le x \le 585, \\ 3.87 \times 10^{-7}(2180 - x) & \text{if } 585 < x < 2180, \\ 0 & \text{otherwise.} \end{cases}$$

Hint: Look at Example 3.9.5.

b. Find the value at risk (VaR) at probability level 0.97 for the portfolio.

13. Prove the special case of Theorem 4.2.5 in which the function g is twice continuously differentiable and $\boldsymbol{X}$ is one-dimensional. You may assume that a twice continuously differentiable convex function has nonnegative second derivative. *Hint:* Expand $g(X)$ around its mean using Taylor's theorem with remainder. Taylor's theorem with remainder says that if $g(x)$ has two continuous derivatives g' and g'' at $x = x_0$, then there exists y between x_0 and x such that

$$g(x) = g(x_0) + (x - x_0)g'(x_0) + \frac{(x - x_0)^2}{2}g''(y).$$

4.3 Variance

Although the mean of a distribution is a useful summary, it does not convey very much information about the distribution. For example, a random variable X with mean 2 has the same mean as the constant random variable Y such that $\Pr(Y = 2) = 1$ even if X is not constant. To distinguish the distribution of X from the distribution of Y in this case, it might be useful to give some measure of how spread out the distribution of X is. The variance of X is one such measure. The standard deviation of X is the square root of the variance. The variance also plays an important role in the approximation methods that arise in Chapter 6.

Example 4.3.1

Stock Price Changes. Consider the prices A and B of two stocks at a time one month in the future. Assume that A has the uniform distribution on the interval [25, 35] and B has the uniform distribution on the interval [15, 45]. It is easy to see (from Exercise 1 in Sec. 4.1) that both stocks have a mean price of 30. But the distributions are very different. For example, A will surely be worth at least 25 while $\Pr(B < 25) = 1/3$. But B has more upside potential also. The p.d.f.'s of these two random variables are plotted in Fig. 4.5. ◀

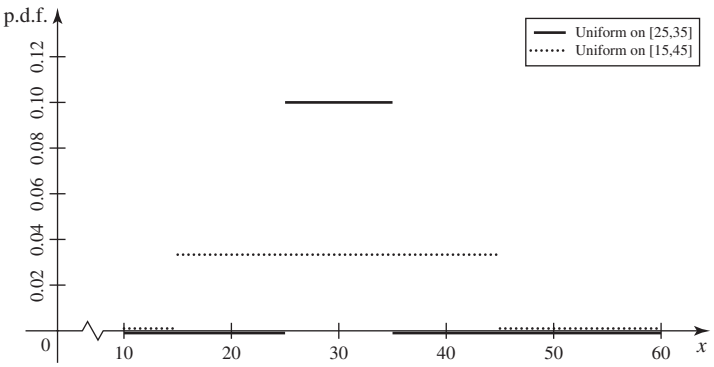

Figure 4.5 The p.d.f.'s of two uniform distributions in Example 4.3.1. Both distributions have mean equal to 30, but they are spread out differently.

Definitions of the Variance and the Standard Deviation

Although the two random prices in Example 4.3.1 have the same mean, price B is more spread out than price A, and it would be good to have a summary of the distribution that makes this easy to see.

Definition 4.3.1 Variance/Standard Deviation. Let X be a random variable with finite mean $\mu = E(X)$. The *variance of* X, denoted by $\text{Var}(X)$, is defined as follows:

$$\text{Var}(X) = E[(X - \mu)^2]. \tag{4.3.1}$$

If X has infinite mean or if the mean of X does not exist, we say that $\text{Var}(X)$ *does not exist*. The *standard deviation of* X is the nonnegative square root of $\text{Var}(X)$ if the variance exists.

If the expectation in Eq. (4.3.1) is infinite, we say that $\text{Var}(X)$ and the standard deviation of X are infinite.

When only one random variable is being discussed, it is common to denote its standard deviation by the symbol σ, and the variance is denoted by σ^2. When more than one random variable is being discussed, the name of the random variable is included as a subscript to the symbol σ, e.g., σ_X would be the standard deviation of X while σ_Y^2 would be the variance of Y.

Example 4.3.2 Stock Price Changes. Return to the two random variables A and B in Example 4.3.1. Using Theorem 4.1.1, we can compute

$$\text{Var}(A) = \int_{25}^{35} (a - 30)^2 \frac{1}{10} da = \frac{1}{10} \int_{-5}^{5} x^2 dx = \frac{1}{10} \frac{x^3}{3} \Bigg|_{x=-5}^{5} = \frac{25}{3},$$

$$\text{Var}(B) = \int_{15}^{45} (b - 30)^2 \frac{1}{30} db = \frac{1}{30} \int_{-15}^{15} y^2 dy = \frac{1}{30} \frac{y^3}{3} \Bigg|_{y=-15}^{15} = 75.$$

So, $\text{Var}(B)$ is nine times as large as $\text{Var}(A)$. The standard deviations of A and B are $\sigma_A = 2.87$ and $\sigma_B = 8.66$. ◀

Note: Variance Depends Only on the Distribution. The variance and standard deviation of a random variable X depend only on the distribution of X, just as the expectation of X depends only on the distribution. Indeed, everything that can be computed from the p.f. or p.d.f. depends only on the distribution. Two random

variables with the same distribution will have the same variance, even if they have nothing to do with each other.

Example 4.3.3 Variance and Standard Deviation of a Discrete Distribution. Suppose that a random variable X can take each of the five values -2, 0, 1, 3, and 4 with equal probability. We shall determine the variance and standard deviation of X.

In this example,

$$E(X) = \frac{1}{5}(-2 + 0 + 1 + 3 + 4) = 1.2.$$

Let $\mu = E(X) = 1.2$, and define $W = (X - \mu)^2$. Then $\text{Var}(X) = E(W)$. We can easily compute the p.f. f of W:

x	-2	0	1	3	4
w	10.24	1.44	0.04	3.24	7.84
$f(w)$	1/5	1/5	1/5	1/5	1/5

It follows that

$$\text{Var}(X) = E(W) = \frac{1}{5}[10.24 + 1.44 + 0.04 + 3.24 + 7.84] = 4.56.$$

The standard deviation of X is the square root of the variance, namely, 2.135. ◀

There is an alternative method for calculating the variance of a distribution, which is often easier to use.

Theorem 4.3.1 Alternative Method for Calculating the Variance. For every random variable X, $\text{Var}(X) = E(X^2) - [E(X)]^2$.

Proof Let $E(X) = \mu$. Then

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

■

Example 4.3.4 Variance of a Discrete Distribution. Once again, consider the random variable X in Example 4.3.3, which takes each of the five values -2, 0, 1, 3, and 4 with equal probability. We shall use Theorem 4.3.1 to compute $\text{Var}(X)$. In Example 4.3.3, we computed the mean of X as $\mu = 1.2$. To use Theorem 4.3.1, we need

$$E(X^2) = \frac{1}{5}[(-2)^2 + 0^2 + 1^2 + 3^2 + 4^2] = 6.$$

Because $E(X) = 1.2$, Theorem 4.3.1 says that

$$\text{Var}(X) = 6 - (1.2)^2 = 4.56,$$

which agrees with the calculation in Example 4.3.3. ◀

The variance (as well as the standard deviation) of a distribution provides a measure of the spread or dispersion of the distribution around its mean μ. A small value of the variance indicates that the probability distribution is tightly concentrated around

μ; a large value of the variance typically indicates that the probability distribution has a wide spread around μ. However, the variance of a distribution, as well as its mean, can be made arbitrarily large by placing even a very small but positive amount of probability far enough from the origin on the real line.

Example 4.3.5 Slight Modification of a Bernoulli Distribution. Let X be a discrete random variable with the following p.d.f.:

$$f(x) = \begin{cases} 0.5 & \text{if } x = 0, \\ 0.499 & \text{if } x = 1, \\ 0.001 & \text{if } x = 10{,}000, \\ 0 & \text{otherwise.} \end{cases}$$

There is a sense in which the distribution of X differs very little from the Bernoulli distribution with parameter 0.5. However, the mean and variance of X are quite different from the mean and variance of the Bernoulli distribution with parameter 0.5. Let Y have the Bernoulli distribution with parameter 0.5. In Example 4.1.3, we computed the mean of Y as $E(Y) = 0.5$. Since $Y^2 = Y$, $E(Y^2) = E(Y) = 0.5$, so $\text{Var}(Y) = 0.5 - 0.5^2 = 0.25$. The means of X and X^2 are also straightforward calculations:

$$E(X) = 0.5 \times 0 + 0.499 \times 1 + 0.001 \times 10{,}000 = 10.499$$

$$E(X^2) = 0.5 \times 0^2 + 0.499 \times 1^2 + 0.001 \times 10{,}000^2 = 100{,}000.499.$$

So $\text{Var}(X) = 99{,}890.27$. The mean and variance of X are much larger than the mean and variance of Y. ◀

Properties of the Variance

We shall now present several theorems that state basic properties of the variance. In these theorems we shall assume that the variances of all the random variables exist. The first theorem concerns the possible values of the variance.

Theorem 4.3.2 For each X, $\text{Var}(X) \geq 0$. If X is a bounded random variable, then $\text{Var}(X)$ must exist and be finite.

Proof Because $\text{Var}(X)$ is the mean of a nonnegative random variable $(X - \mu)^2$, it must be nonnegative according to Theorem 4.2.2. If X is bounded, then the mean exists, and hence the variance exists. Furthermore, if X is bounded the so too is $(X - \mu)^2$, so the variance must be finite. ■

The next theorem shows that the variance of a random variable X cannot be 0 unless the entire probability distribution of X is concentrated at a single point.

Theorem 4.3.3 $\text{Var}(X) = 0$ if and only if there exists a constant c such that $\Pr(X = c) = 1$.

Proof Suppose first that there exists a constant c such that $\Pr(X = c) = 1$. Then $E(X) = c$, and $\Pr[(X - c)^2 = 0] = 1$. Therefore,

$$\text{Var}(X) = E[(X - c)^2] = 0.$$

Conversely, suppose that $\text{Var}(X) = 0$. Then $\Pr[(X - \mu)^2 \geq 0] = 1$ but $E[(X - \mu)^2] = 0$. It follows from Theorem 4.2.3 that

$$\Pr[(X - \mu)^2 = 0] = 1.$$

Hence, $\Pr(X = \mu) = 1$. ■

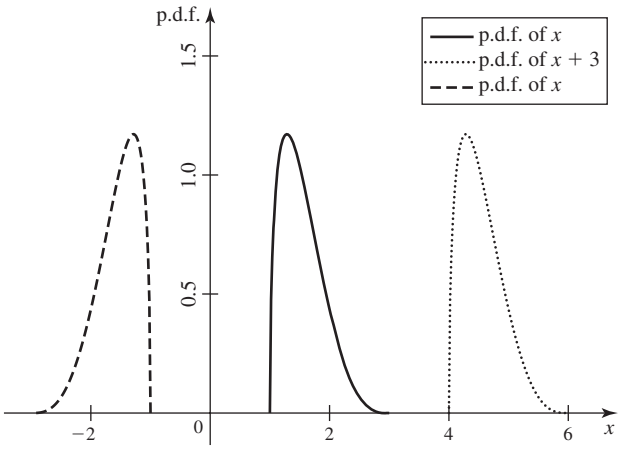

Figure 4.6 The p.d.f. of a random variable X together with the p.d.f.'s of $X + 3$ and $-X$. Note that the spreads of all three distributions appear the same.

Theorem 4.3.4 For constants a and b, let $Y = aX + b$. Then

$$\text{Var}(Y) = a^2 \text{Var}(X),$$

and $\sigma_Y = |a|\sigma_X$.

Proof If $E(X) = \mu$, then $E(Y) = a\mu + b$ by Theorem 4.2.1. Therefore,

$$\text{Var}(Y) = E[(aX + b - a\mu - b)^2] = E[(aX - a\mu)^2]$$
$$= a^2 E[(X - \mu)^2] = a^2 \text{Var}(X).$$

Taking the square root of $\text{Var}(Y)$ yields $|a|\sigma_X$. ■

It follows from Theorem 4.3.4 that $\text{Var}(X + b) = \text{Var}(X)$ for every constant b. This result is intuitively plausible, since shifting the entire distribution of X a distance of b units along the real line will change the mean of the distribution by b units but the shift will not affect the dispersion of the distribution around its mean. Figure 4.6 shows the p.d.f. a random variable X together with the p.d.f. of $X + 3$ to illustrate how a shift of the distribution does not affect the spread.

Similarly, it follows from Theorem 4.3.4 that $\text{Var}(-X) = \text{Var}(X)$. This result also is intuitively plausible, since reflecting the entire distribution of X with respect to the origin of the real line will result in a new distribution that is the mirror image of the original one. The mean will be changed from μ to $-\mu$, but the total dispersion of the distribution around its mean will not be affected. Figure 4.6 shows the p.d.f. of a random variable X together with the p.d.f. of $-X$ to illustrate how a reflection of the distribution does not affect the spread.

Example 4.3.6 Calculating the Variance and Standard Deviation of a Linear Function. Consider the same random variable X as in Example 4.3.3, which takes each of the five values -2, 0, 1, 3, and 4 with equal probability. We shall determine the variance and standard deviation of $Y = 4X - 7$.

In Example 4.3.3, we computed the mean of X as $\mu = 1.2$ and the variance as 4.56. By Theorem 4.3.4,

$$\text{Var}(Y) = 16 \text{Var}(X) = 72.96.$$

Also, the standard deviation σ of Y is

$$\sigma_Y = 4\sigma_X = 4(4.56)^{1/2} = 8.54.$$ ◀

The next theorem provides an alternative method for calculating the variance of a sum of independent random variables.

Theorem 4.3.5 If $X_1, \ldots, X_n$ are independent random variables with finite means, then

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

Proof Suppose first that $n = 2$. If $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$, then

$$E(X_1 + X_2) = \mu_1 + \mu_2.$$

Therefore,

$$\begin{aligned}\text{Var}(X_1 + X_2) &= E[(X_1 + X_2 - \mu_1 - \mu_2)^2] \\ &= E[(X_1 - \mu_1)^2 + (X_2 - \mu_2)^2 + 2(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \text{Var}(X_1) + \text{Var}(X_2) + 2E[(X_1 - \mu_1)(X_2 - \mu_2)].\end{aligned}$$

Since X_1 and X_2 are independent,

$$\begin{aligned}E[(X_1 - \mu_1)(X_2 - \mu_2)] &= E(X_1 - \mu_1)E(X_2 - \mu_2) \\ &= (\mu_1 - \mu_1)(\mu_2 - \mu_2) \\ &= 0.\end{aligned}$$

It follows, therefore, that

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2).$$

The theorem can now be established for each positive integer n by an induction argument. ■

It should be emphasized that the random variables in Theorem 4.3.5 must be independent. The variance of the sum of random variables that are not independent will be discussed in Sec. 4.6. By combining Theorems 4.3.4 and 4.3.5, we can now obtain the following corollary.

Corollary 4.3.1 If $X_1, \ldots, X_n$ are independent random variables with finite means, and if $a_1, \ldots, a_n$ and b are arbitrary constants, then

$$\text{Var}(a_1X_1 + \cdots + a_nX_n + b) = a_1^2 \text{Var}(X_1) + \cdots + a_n^2 \text{Var}(X_n).$$ ■

Example 4.3.7 Investment Portfolio. An investor with \$100,000 to invest wishes to construct a portfolio consisting of shares of one or both of two available stocks and possibly some fixed-rate investments. Suppose that the two stocks have random rates of return R_1 and R_2 per share for a period of one year. Suppose that R_1 has a distribution with mean 6 and variance 55, while R_2 has mean 4 and variance 28. Suppose that the first stock costs \$60 per share and the second costs \$48 per share. Suppose that money can also be invested at a fixed rate of 3.6 percent per year. The portfolio will consist of s_1 shares of the first stock, s_2 shares of the second stock, and all remaining money ($\$s_3$) invested at the fixed rate. The return on this portfolio will be

$$s_1R_1 + s_2R_2 + 0.036s_3,$$

where the coefficients are constrained by

$$60s_1 + 48s_2 + s_3 = 100{,}000, \tag{4.3.2}$$

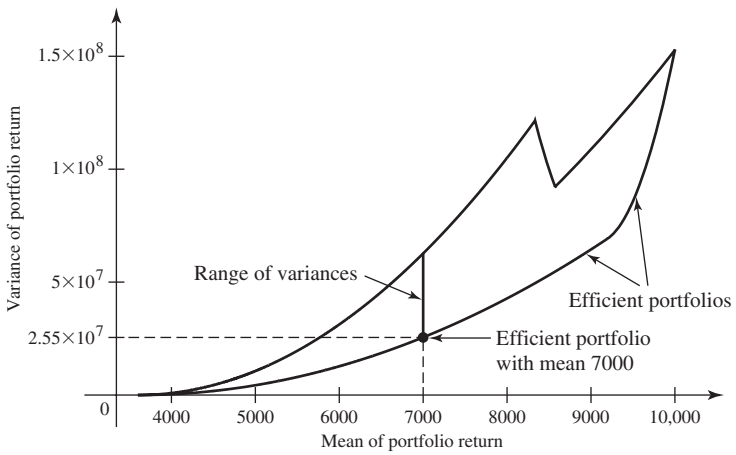

Figure 4.7 The set of all means and variances of investment portfolios in Example 4.3.7. The solid vertical line shows the range of possible variances for portfoloios with a mean of 7000.

as well as $s_1, s_2, s_3 \geq 0$. For now, we shall assume that R_1 and R_2 are independent. The mean and the variance of the return on the portfolio will be

$$E(s_1R_1 + s_2R_2 + 0.036s_3) = 6s_1 + 4s_2 + 0.036s_3,$$
$$\text{Var}(s_1R_1 + s_2R_2 + 0.036s_3) = 55s_1^2 + 28s_2^2.$$

One method for comparing a class of portfolios is to say that portfolio A is at least as good as portfolio B if the mean return for A is at least as large as the mean return for B and if the variance for A is no larger than the variance of B. (See Markowitz, 1987, for a classic treatment of such methods.) The reason for preferring smaller variance is that large variance is associated with large deviations from the mean, and for portfolios with a common mean, some of the large deviations are going to have to be below the mean, leading to the risk of large losses. Figure 4.7 is a plot of the pairs (mean, variance) for all of the possible portfolios in this example. That is, for each (s_1, s_2, s_3) that satisfy (4.3.2), there is a point in the outlined region of Fig. 4.7. The points to the right and toward the bottom are those that have the largest mean return for a fixed variance, and the ones that have the smallest variance for a fixed mean return. These portfolios are called *efficient*. For example, suppose that the investor would like a mean return of 7000. The vertical line segment above 7000 on the horizontal axis in Fig. 4.7 indicates the possible variances of all portfolios with mean return of 7000. Among these, the portfolio with the smallest variance is efficient and is indicated in Fig. 4.7. This portfolio has $s_1 = 524.7$, $s_2 = 609.7$, $s_3 = 39{,}250$, and variance 2.55×10^7. So, every portfolio with mean return greater than 7000 must have variance larger than 2.55×10^7, and every portfolio with variance less than 2.55×10^7 must have mean return smaller than 7000. ◀

The Variance of a Binomial Distribution

We shall now consider again the method of generating a binomial distribution presented in Sec. 4.2. Suppose that a box contains red balls and blue balls, and that the proportion of red balls is p $(0 \leq p \leq 1)$. Suppose also that a random sample of n balls is selected from the box with replacement. For $i = 1, \ldots, n$, let $X_i = 1$ if the ith ball that is selected is red, and let $X_i = 0$ otherwise. If X denotes the total number of red balls in the sample, then $X = X_1 + \cdots + X_n$ and X will have the binomial distribution with parameters n and p.

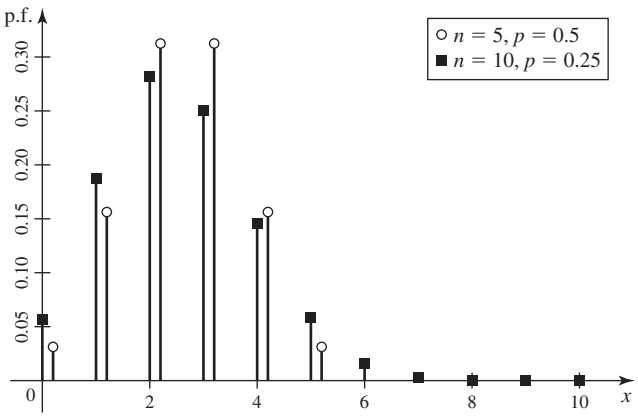

Figure 4.8 Two binomial distributions with the same mean (2.5) but different variances.

Since $X_1, \ldots, X_n$ are independent, it follows from Theorem 4.3.5 that

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i).$$

According to Example 4.1.3, $E(X_i) = p$ for $i = 1, \ldots, n$. Since $X_i^2 = X_i$ for each i, $E(X_i^2) = E(X_i) = p$. Therefore, by Theorem 4.3.1,

$$\begin{aligned}\text{Var}(X_i) &= E(X_i^2) - [E(X_i)]^2 \\ &= p - p^2 = p(1-p).\end{aligned}$$

It now follows that

$$\text{Var}(X) = np(1-p). \tag{4.3.3}$$

Figure 4.8 compares two different binomial distributions with the same mean (2.5) but different variances (1.25 and 1.875). One can see how the p.f. of the distribution with the larger variance ($n = 10$, $p = 0.25$) is higher at more extreme values and lower at more central values than is the p.f. of the distribution with the smaller variance ($n = 5$, $p = 0.5$). Similarly, Fig. 4.5 compares two uniform distributions with the same mean (30) and different variances (8.33 and 75). The same pattern appears, namely that the distribution with larger variance has higher p.d.f. at more extreme values and lower p.d.f. at more central values.

Interquartile Range

Example 4.3.8

The Cauchy Distribution. In Example 4.1.8, we saw a distribution (the Cauchy distribution) whose mean did not exist, and hence its variance does not exist. But, we might still want to describe how spread out such a distribution is. For example, if X has the Cauchy distribution and $Y = 2X$, it stands to reason that Y is twice as spread out as X is, but how do we quantify this? ◀

There is a measure of spread that exists for every distribution, regardless of whether or not the distribution has a mean or variance. Recall from Definition 3.3.2 that the quantile function for a random variable is the inverse of the c.d.f., and it is defined for every random variable.

Definition 4.3.2 Interquartile Range (IQR). Let X be a random variable with quantile function $F^{-1}(p)$ for $0 < p < 1$. The *interquartile range (IQR)* is defined to be $F^{-1}(0.75) - F^{-1}(0.25)$.

In words, the IQR is the length of the interval that contains the middle half of the distribution.

Example 4.3.9 The Cauchy Distribution. Let X have the Cauchy distribution. The c.d.f. F of X can be found using a trigonometric substitution in the following integral:

$$F(x) = \int_{-\infty}^{x} \frac{dy}{\pi(1+y^2)} = \frac{1}{2} + \frac{\tan^{-1}(x)}{\pi},$$

where $\tan^{-1}(x)$ is the principal inverse of the tangent function, taking values from $-\pi/2$ to $\pi/2$ as x runs from $-\infty$ to ∞. The quantile function of X is then $F^{-1}(p) = \tan[\pi(p - 1/2)]$ for $0 < p < 1$. The IQR is

$$F^{-1}(0.75) - F^{-1}(0.25) = \tan(\pi/4) - \tan(-\pi/4) = 2.$$

It is not difficult to show that, if $Y = 2X$, then the IQR of Y is 4. (See Exercise 14.) ◀

Summary

The variance of X, denoted by $\mathrm{Var}(X)$, is the mean of $[X - E(X)]^2$ and measures how spread out the distribution of X is. The variance also equals $E(X^2) - [E(X)]^2$. The standard deviation is the square root of the variance. The variance of $aX + b$, where a and b are constants, is $a^2\,\mathrm{Var}(X)$. The variance of the sum of independent random variables is the sum of the variances. As an example, the variance of the binomial distribution with parameters n and p is $np(1 - p)$. The interquartile range (IQR) is the difference between the 0.75 and 0.25 quantiles. The IQR is a measure of spread that exists for every distribution.

Exercises

1. Suppose that X has the uniform distribution on the interval $[0, 1]$. Compute the variance of X.

2. Suppose that one word is selected at random from the sentence THE GIRL PUT ON HER BEAUTIFUL RED HAT. If X denotes the number of letters in the word that is selected, what is the value of $\mathrm{Var}(X)$?

3. For all numbers a and b such that $a < b$, find the variance of the uniform distribution on the interval $[a, b]$.

4. Suppose that X is a random variable for which $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Show that $E[X(X - 1)] = \mu(\mu - 1) + \sigma^2$.

5. Let X be a random variable for which $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, and let c be an arbitrary constant. Show that

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

6. Suppose that X and Y are independent random variables whose variances exist and such that $E(X) = E(Y)$. Show that

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

7. Suppose that X and Y are independent random variables for which $\mathrm{Var}(X) = \mathrm{Var}(Y) = 3$. Find the values of **(a)** $\mathrm{Var}(X - Y)$ and **(b)** $\mathrm{Var}(2X - 3Y + 1)$.

8. Construct an example of a distribution for which the mean is finite but the variance is infinite.

9. Let X have the discrete uniform distribution on the integers $1, \ldots, n$. Compute the variance of X. *Hint:* You may wish to use the formula $\sum_{k=1}^{n} k^2 = n(n + 1) \cdot (2n + 1)/6$.

10. Consider the example efficient portfolio at the end of Example 4.3.7. Suppose that R_i has the uniform distribution on the interval $[a_i, b_i]$ for $i = 1, 2$.

a. Find the two intervals $[a_1, b_1]$ and $[a_2, b_2]$. *Hint:* The intervals are determined by the means and variances.

b. Find the value at risk (VaR) for the example portfolio at probability level 0.97. *Hint:* Review Example 3.9.5 to see how to find the p.d.f. of the sum of two uniform random variables.

11. Let X have the uniform distribution on the interval $[0, 1]$. Find the IQR of X.

12. Let X have the p.d.f. $f(x) = \exp(-x)$ for $x \geq 0$, and $f(x) = 0$ for $x < 0$. Find the IQR of X.

13. Let X have the binomial distribution with parameters 5 and 0.3. Find the IQR of X. *Hint:* Return to Example 3.3.9 and Table 3.1.

14. Let X be a random variable whose interquartile range is η. Let $Y = 2X$. Prove that the interquartile range of Y is 2η.

4.4 Moments

For a random variable X, the means of powers X^k (called moments) for $k > 2$ have useful theoretical properties, and some of them are used for additional summaries of a distribution. The moment generating function is a related tool that aids in deriving distributions of sums of independent random variables and limiting properties of distributions.

Existence of Moments

For each random variable X and every positive integer k, the expectation $E(X^k)$ is called the kth *moment of* X. In particular, in accordance with this terminology, the mean of X is the first moment of X.

It is said that the kth moment exists if and only if $E(|X|^k) < \infty$. If the random variable X is bounded, that is, if there are finite numbers a and b such that $\Pr(a \leq X \leq b) = 1$, then all moments of X must necessarily exist. It is possible, however, that all moments of X exist even though X is not bounded. It is shown in the next theorem that if the kth moment of X exists, then all moments of lower order must also exist.

Theorem 4.4.1 If $E(|X|^k) < \infty$ for some positive integer k, then $E(|X|^j) < \infty$ for every positive integer j such that $j < k$.

Proof We shall assume, for convenience, that the distribution of X is continuous and the p.d.f. is f. Then

$$\begin{aligned} E(|X|^j) &= \int_{-\infty}^{\infty} |x|^j f(x)\, dx \\ &= \int_{|x|\leq 1} |x|^j f(x)\, dx + \int_{|x|>1} |x|^j f(x)\, dx \\ &\leq \int_{|x|\leq 1} 1 \cdot f(x)\, dx + \int_{|x|>1} |x|^k f(x)\, dx \\ &\leq \Pr(|X| \leq 1) + E(|X|^k). \end{aligned}$$

By hypothesis, $E(|X|^k) < \infty$. It therefore follows that $E(|X|^j) < \infty$. A similar proof holds for a discrete or a more general type of distribution. ■

In particular, it follows from Theorem 4.4.1 that if $E(X^2) < \infty$, then both the mean of X and the variance of X exist. Theorem 4.4.1 extends to the case in which

j and k are arbitrary positive numbers rather than just integers. (See Exercise 15 in this section.) We will not make use of such a result in this text, however.

Central Moments Suppose that X is a random variable for which $E(X) = \mu$. For every positive integer k, the expectation $E[(X - \mu)^k]$ is called the kth *central moment of* X or the kth *moment of* X *about the mean*. In particular, in accordance with this terminology, the variance of X is the second central moment of X.

For every distribution, the first central moment must be 0 because

$$E(X - \mu) = \mu - \mu = 0.$$

Furthermore, if the distribution of X is symmetric with respect to its mean μ, and if the central moment $E[(X - \mu)^k]$ exists for a given odd integer k, then the value of $E[(X - \mu)^k]$ will be 0 because the positive and negative terms in this expectation will cancel one another.

Example 4.4.1 **A Symmetric p.d.f.** Suppose that X has a continuous distribution for which the p.d.f. has the following form:

$$f(x) = ce^{-(x-3)^2/2} \quad \text{for } -\infty < x < \infty.$$

We shall determine the mean of X and all the central moments.

It can be shown that for every positive integer k,

$$\int_{-\infty}^{\infty} |x|^k e^{-(x-3)^2/2}\, dx < \infty.$$

Hence, all the moments of X exist. Furthermore, since $f(x)$ is symmetric with respect to the point $x = 3$, then $E(X) = 3$. Because of this symmetry, it also follows that $E[(X - 3)^k] = 0$ for every odd positive integer k. For even $k = 2n$, we can find a recursive formula for the sequence of central moments. First, let $y = x - \mu$ in all the integral fomulas. Then, for $n \geq 1$, the $2n$th central moment is

$$m_{2n} = \int_{-\infty}^{\infty} y^{2n} c e^{-y^2/2} dy.$$

Use integration by parts with $u = y^{2n-1}$ and $dv = ye^{-y^2/2}dy$. It follows that $du = (2n - 1)y^{2n-2}dy$ and $v = -e^{-y^2/2}$. So,

$$\begin{aligned} m_{2n} &= \int_{-\infty}^{\infty} u\,dv = uv|_{y=-\infty}^{\infty} - \int_{-\infty}^{\infty} v\,du \\ &= -y^{2n-1}e^{-y^2/2}\Big|_{y=-\infty}^{\infty} + (2n-1)\int_{-\infty}^{\infty} y^{2n-2} c e^{-y^2/2} dy \\ &= (2n-1)m_{2(n-1)}. \end{aligned}$$

Because $y^0 = 1$, m_0 is just the integral of the p.d.f.; hence, $m_0 = 1$. It follows that $m_{2n} = \prod_{i=1}^{n}(2i - 1)$ for $n = 1, 2, \ldots$. So, for example, $m_2 = 1$, $m_4 = 3$, $m_6 = 15$, and so on. ◀

Skewness In Example 4.4.1, we saw that the odd central moments are all 0 for a distribution that is symmetric. This leads to the following distributional summary that is used to measure lack of symmetry.

Definition 4.4.1 **Skewness.** Let X be a random variable with mean μ, standard deviation σ, and finite third moment. The *skewness* of X is defined to be $E[(X - \mu)^3]/\sigma^3$.

The reason for dividing the third central moment by σ^3 is to make the skewness measure only the lack of symmetry rather than the spread of the distribution.

Example 4.4.2 Skewness of Binomial Distributions. Let X have the binomial distribution with parameters 10 and 0.25. The p.f. of this distribution appears in Fig. 4.8. It is not difficult to see that the p.f. is not symmetric. The skewness can be computed as follows: First, note that the mean is $\mu = 10 \times 0.25 = 2.5$ and that the standard deviation is

$$\sigma = (10 \times 0.25 \times 0.75)^{1/2} = 1.369.$$

Second, compute

$$E[(X-2.5)^3] = (0-2.5)^3\binom{10}{0}0.25^0\,0.75^{10} + \cdots + (10-2.5)^3\binom{10}{10}0.25^{00}\,0.75^0$$
$$= 0.9375.$$

Finally, the skewness is

$$\frac{0.9375}{1.369^3} = 0.3652.$$

For comparison, the skewness of the binomial distribution with parameters 10 and 0.2 is 0.4743, and the skewness of the binomial distribution with parameters 10 and 0.3 is 0.2761. The absolute value of the skewness increases as the probability of success moves away from 0.5. It is straightforward to show that the skewness of the binomial distribution with parameters n and p is the negative of the skewness of the binomial distribution with parameters n and $1-p$. (See Exercise 16 in this section.) ◀

Moment Generating Functions

We shall now consider a different way to characterize the distribution of a random variable that is more closely related to its moments than to where its probability is distributed.

Definition 4.4.2 Moment Generating Function. Let X be a random variable. For each real number t, define

$$\psi(t) = E(e^{tX}). \tag{4.4.1}$$

The function $\psi(t)$ is called the *moment generating function* (abbreviated m.g.f.) of X.

Note: The Moment Generating Function of X Depends Only on the Distribution of X. Since the m.g.f. is the expected value of a function of X, it must depend only on the distribution of X. If X and Y have the same distribution, they must have the same m.g.f.

If the random variable X is bounded, then the expectation in Eq. (4.4.1) must be finite for all values of t. In this case, therefore, the m.g.f. of X will be finite for all values of t. On the other hand, if X is not bounded, then the m.g.f. might be finite for some values of t and might not be finite for others. It can be seen from Eq. (4.4.1), however, that for every random variable X, the m.g.f. $\psi(t)$ must be finite at the point $t = 0$ and at that point its value must be $\psi(0) = E(1) = 1$.

The next result explains how the name "moment generating function" arose.

Theorem 4.4.2 Let X be a random variables whose m.g.f. $\psi(t)$ is finite for all values of t in some open interval around the point $t = 0$. Then, for each integer $n > 0$, the nth moment of X,

$E(X^n)$, is finite and equals the nth derivative $\psi^{(n)}(t)$ at $t = 0$. That is, $E(X^n) = \psi^{(n)}(0)$ for $n = 1, 2, \ldots$.

We sketch the proof at the end of this section.

Example 4.4.3 Calculating an m.g.f. Suppose that X is a random variable for which the p.d.f. is as follows:

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

We shall determine the m.g.f. of X and also $\text{Var}(X)$.

For each real number t,

$$\psi(t) = E(e^{tX}) = \int_0^\infty e^{tx} e^{-x}\, dx$$
$$= \int_0^\infty e^{(t-1)x}\, dx.$$

The final integral in this equation will be finite if and only if $t < 1$. Therefore, $\psi(t)$ is finite only for $t < 1$. For each such value of t,

$$\psi(t) = \frac{1}{1-t}.$$

Since $\psi(t)$ is finite for all values of t in an open interval around the point $t = 0$, all moments of X exist. The first two derivatives of ψ are

$$\psi'(t) = \frac{1}{(1-t)^2} \quad \text{and} \quad \psi''(t) = \frac{2}{(1-t)^3}.$$

Therefore, $E(X) = \psi'(0) = 1$ and $E(X^2) = \psi''(0) = 2$. It now follows that

$$\text{Var}(X) = \psi''(0) - [\psi'(0)]^2 = 1. \quad \blacktriangleleft$$

Properties of Moment Generating Functions

We shall now present three basic theorems pertaining to moment generating functions.

Theorem 4.4.3 Let X be a random variable for which the m.g.f. is ψ_1; let $Y = aX + b$, where a and b are given constants; and let ψ_2 denote the m.g.f. of Y. Then for every value of t such that $\psi_1(at)$ is finite,

$$\psi_2(t) = e^{bt}\psi_1(at). \tag{4.4.2}$$

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

$$\psi_2(t) = E(e^{tY}) = E[e^{t(aX+b)}] = e^{bt}E(e^{atX}) = e^{bt}\psi_1(at). \quad \blacksquare$$

Example 4.4.4 Calculating the m.g.f. of a Linear Function. Suppose that the distribution of X is as specified in Example 4.4.3. We saw that the m.g.f. of X for $t < 1$ is

$$\psi_1(t) = \frac{1}{1-t}.$$

If $Y = 3 - 2X$, then the m.g.f. of Y is finite for $t > -1/2$ and will have the value

$$\psi_2(t) = e^{3t}\psi_1(-2t) = \frac{e^{3t}}{1+2t}. \quad \blacktriangleleft$$

The next theorem shows that the m.g.f. of the sum of an arbitrary number of independent random variables has a very simple form. Because of this property, the m.g.f. is an important tool in the study of such sums.

Theorem 4.4.4 Suppose that $X_1, \ldots, X_n$ are n independent random variables; and for $i = 1, \ldots, n$, let ψ_i denote the m.g.f. of X_i. Let $Y = X_1 + \cdots + X_n$, and let the m.g.f. of Y be denoted by ψ. Then for every value of t such that $\psi_i(t)$ is finite for $i = 1, \ldots, n$,

$$\psi(t) = \prod_{i=1}^{n} \psi_i(t). \tag{4.4.3}$$

Proof By definition,

$$\psi(t) = E(e^{tY}) = E[e^{t(X_1+\cdots+X_n)}] = E\left(\prod_{i=1}^{n} e^{tX_i}\right).$$

Since the random variables $X_1, \ldots, X_n$ are independent, it follows from Theorem 4.2.6 that

$$E\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} E(e^{tX_i}).$$

Hence,

$$\psi(t) = \prod_{i=1}^{n} \psi_i(t). \qquad \blacksquare$$

The Moment Generating Function for the Binomial Distribution Suppose that a random variable X has the binomial distribution with parameters n and p. In Sections 4.2 and 4.3, the mean and the variance of X were determined by representing X as the sum of n independent random variables $X_1, \ldots, X_n$. In this representation, the distribution of each variable X_i is as follows:

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

We shall now use this representation to determine the m.g.f. of $X = X_1 + \cdots + X_n$.

Since each of the random variables $X_1, \ldots, X_n$ has the same distribution, the m.g.f. of each variable will be the same. For $i = 1, \ldots, n$, the m.g.f. of X_i is

$$\begin{aligned}\psi_i(t) = E(e^{tX_i}) &= (e^t)\Pr(X_i = 1) + (1)\Pr(X_i = 0)\\ &= pe^t + 1 - p.\end{aligned}$$

It follows from Theorem 4.4.4 that the m.g.f. of X in this case is

$$\psi(t) = (pe^t + 1 - p)^n. \tag{4.4.4}$$

Uniqueness of Moment Generating Functions We shall now state one more important property of the m.g.f. The proof of this property is beyond the scope of this book and is omitted.

Theorem 4.4.5 If the m.g.f.'s of two random variables X_1 and X_2 are finite and identical for all values of t in an open interval around the point $t = 0$, then the probability distributions of X_1 and X_2 must be identical. $\blacksquare$

Theorem 4.4.5 is the justification for the claim made at the start of this discussion, namely, that the m.g.f. is another way to characterize the distribution of a random variable.

The Additive Property of the Binomial Distribution Moment generating functions provide a simple way to derive the distribution of the sum of two independent binomial random variables with the same second parameter.

Theorem 4.4.6 If X_1 and X_2 are independent random variables, and if X_i has the binomial distribution with parameters n_i and p $(i = 1, 2)$, then $X_1 + X_2$ has the binomial distribution with parameters $n_1 + n_2$ and p.

Proof L et ψ_i denote the m.g.f. of X_i for $i = 1, 2$. It follows from Eq. (4.4.4) that

$$\psi_i(t) = (pe^t + 1 - p)^{n_i}.$$

Let ψ denote the m.g.f. of $X_1 + X_2$. Then, by Theorem 4.4.4,

$$\psi(t) = (pe^t + 1 - p)^{n_1+n_2}.$$

It can be seen from Eq. (4.4.4) that this function ψ is the m.g.f. of the binomial distribution with parameters $n_1 + n_2$ and p. Hence, by Theorem 4.4.5, the distribution of $X_1 + X_2$ must be that binomial distribution. ■

Sketch of the Proof of Theorem 4.4.2

First, we indicate why all moments of X are finite. Let $t > 0$ be such that both $\psi(t)$ and $\psi(-t)$ are finite. Define $g(x) = e^{tx} + e^{-tx}$. Notice that

$$E[g(X)] = \psi(t) + \psi(-t) < \infty. \tag{4.4.5}$$

On every bounded interval of x values, $g(x)$ is bounded. For each integer $n > 0$, as $|x| \to \infty$, $g(x)$ is eventually larger than $|x|^n$. It follows from these facts and (4.4.5) that $E|X^n| < \infty$.

Although it is beyond the scope of this book, it can be shown that the derivative $\psi'(t)$ exists at the point $t = 0$, and that at $t = 0$, the derivative of the expectation in Eq. (4.4.1) must be equal to the expectation of the derivative. Thus,

$$\psi'(0) = \left[\frac{d}{dt}E(e^{tX})\right]_{t=0} = E\left[\left(\frac{d}{dt}e^{tX}\right)_{t=0}\right].$$

But

$$\left(\frac{d}{dt}e^{tX}\right)_{t=0} = (Xe^{tX})_{t=0} = X.$$

It follows that

$$\psi'(0) = E(X).$$

In other words, the derivative of the m.g.f. $\psi(t)$ at $t = 0$ is the mean of X.

Furthermore, it can be shown that it is possible to differentiate $\psi(t)$ an arbitrary number of times at the point $t = 0$. For $n = 1, 2, \ldots$, the nth derivative $\psi^{(n)}(0)$ at $t = 0$ will satisfy the following relation:

$$\psi^{(n)}(0) = \left[\frac{d^n}{dt^n}E(e^{tX})\right]_{t=0} = E\left[\left(\frac{d^n}{dt^n}e^{tX}\right)_{t=0}\right]$$
$$= E[(X^n e^{tX})_{t=0}] = E(X^n).$$

Thus, $\psi'(0) = E(X)$, $\psi''(0) = E(X^2)$, $\psi'''(0) = E(X^3)$, and so on. Hence, we see that the m.g.f., if it is finite in an open interval around $t = 0$, can be used to generate all of the moments of the distribution by taking derivatives at $t = 0$.

Summary

If the kth moment of a random variable exists, then so does the jth moment for every $j < k$. The moment generating function of X, $\psi(t) = E(e^{tX})$, if it is finite for t in a neighborhood of 0, can be used to find moments of X. The kth derivative of $\psi(t)$ at $t = 0$ is $E(X^k)$. The m.g.f. characterizes the distribution in the sense that all random variables that have the same m.g.f. have the same distribution.

Exercises

1. If X has the uniform distribution on the interval $[a, b]$, what is the value of the fifth central moment of X?

2. If X has the uniform distribution on the interval $[a, b]$, write a formula for every even central moment of X.

3. Suppose that X is a random variable for which $E(X) = 1$, $E(X^2) = 2$, and $E(X^3) = 5$. Find the value of the third central moment of X.

4. Suppose that X is a random variable such that $E(X^2)$ is finite. **(a)** Show that $E(X^2) \geq [E(X)]^2$. **(b)** Show that $E(X^2) = [E(X)]^2$ if and only if there exists a constant c such that $\Pr(X = c) = 1$. *Hint:* $\mathrm{Var}(X) \geq 0$.

5. Suppose that X is a random variable with mean μ and variance σ^2, and that the fourth moment of X is finite. Show that

$$E[(X - \mu)^4] \geq \sigma^4.$$

6. Suppose that X has the uniform distribution on the interval $[a, b]$. Determine the m.g.f. of X.

7. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = \frac{1}{4}(3e^t + e^{-t}) \quad \text{for } -\infty < t < \infty.$$

Find the mean and the variance of X.

8. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = e^{t^2+3t} \quad \text{for } -\infty < t < \infty.$$

Find the mean and the variance of X.

9. Let X be a random variable with mean μ and variance σ^2, and let $\psi_1(t)$ denote the m.g.f. of X for $-\infty < t < \infty$. Let c be a given positive constant, and let Y be a random variable for which the m.g.f. is

$$\psi_2(t) = e^{c[\psi_1(t)-1]} \quad \text{for } -\infty < t < \infty.$$

Find expressions for the mean and the variance of Y in terms of the mean and the variance of X.

10. Suppose that the random variables X and Y are i.i.d. and that the m.g.f. of each is

$$\psi(t) = e^{t^2+3t} \quad \text{for } -\infty < t < \infty.$$

Find the m.g.f. of $Z = 2X - 3Y + 4$.

11. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = \frac{1}{5}e^t + \frac{2}{5}e^{4t} + \frac{2}{5}e^{8t} \quad \text{for } -\infty < t < \infty.$$

Find the probability distribution of X. *Hint:* It is a simple discrete distribution.

12. Suppose that X is a random variable for which the m.g.f. is as follows:

$$\psi(t) = \frac{1}{6}(4 + e^t + e^{-t}) \quad \text{for } -\infty < t < \infty.$$

Find the probability distribution of X.

13. Let X have the Cauchy distribution (see Example 4.1.8). Prove that the m.g.f. $\psi(t)$ is finite only for $t = 0$.

14. Let X have p.d.f.

$$f(x) = \begin{cases} x^{-2} & \text{if } x > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Prove that the m.g.f. $\psi(t)$ is finite for all $t \leq 0$ but for no $t > 0$.

15. Prove the following extension of Theorem 4.4.1: If $E(|X|^a) < \infty$ for some positive number a, then $E(|X|^b) < \infty$ for every positive number $b < a$. Give the proof for the case in which X has a discrete distribution.

16. Let X have the binomial distribution with parameters n and p. Let Y have the binomial distribution with parameters n and $1 - p$. Prove that the skewness of Y is the negative of the skewness of X. *Hint:* Let $Z = n - X$ and show that Z has the same distribution as Y.

17. Find the skewness of the distribution in Example 4.4.3.

4.5 The Mean and the Median

Although the mean of a distribution is a measure of central location, the median (see Definition 3.3.3) is also a measure of central location for a distribution. This section presents some comparisons and contrasts between these two location summaries of a distribution.

The Median

It was mentioned in Sec. 4.1 that the mean of a probability distribution on the real line will be at the center of gravity of that distribution. In this sense, the mean of a distribution can be regarded as the *center* of the distribution. There is another point on the line that might also be regarded as the center of the distribution. Suppose that there is a point m_0 that divides the total probability into two equal parts, that is, the probability to the left of m_0 is 1/2, and the probability to the right of m_0 is also 1/2. For a continuous distribution, the median of the distribution introduced in Definition 3.3.3 is such a number. If there is such an m_0, it could legitimately be called a center of the distribution. It should be noted, however, that for some discrete distributions there will not be any point at which the total probability is divided into two parts that are exactly equal. Moreover, for other distributions, which may be either discrete or continuous, there will be more than one such point. Therefore, the formal definition of a median, which will now be given, must be general enough to include these possibilities.

Definition 4.5.1 Median. Let X be a random variable. Every number m with the following property is called a *median* of the distribution of X:

$$\Pr(X \le m) \ge 1/2 \quad \text{and} \quad \Pr(X \ge m) \ge 1/2.$$

Another way to understand this definition is that a median is a point m that satisfies the following two requirements: First, if m is included with the values of X to the left of m, then

$$\Pr(X \le m) \ge \Pr(X > m).$$

Second, if m is included with the values of X to the right of m, then

$$\Pr(X \ge m) \ge \Pr(X < m).$$

If there is a number m such that $\Pr(X < m) = \Pr(X > m)$, that is, if the number m does actually divide the total probability into two equal parts, then m will of course be a median of the distribution of X (see Exercise 16).

Note: Multiple Medians. One can prove that every distribution must have at least one median. Indeed, the 1/2 quantile from Definition 3.3.2 is a median. (See Exercise 1.) For some distributions, every number in some interval is a median. In such

cases, the 1/2 quantile is the minimum of the set of all medians. When a whole interval of numbers are medians of a distribution, some writers refer to the midpoint of the interval as the median.

Example 4.5.1 The Median of a Discrete Distribution. Suppose that X has the following discrete distribution:

$$\Pr(X = 1) = 0.1, \qquad \Pr(X = 2) = 0.2,$$
$$\Pr(X = 3) = 0.3, \qquad \Pr(X = 4) = 0.4.$$

The value 3 is a median of this distribution because $\Pr(X \le 3) = 0.6$, which is greater than 1/2, and $\Pr(X \ge 3) = 0.7$, which is also greater than 1/2. Furthermore, 3 is the unique median of this distribution. ◀

Example 4.5.2 A Discrete Distribution for Which the Median Is Not Unique. Suppose that X has the following discrete distribution:

$$\Pr(X = 1) = 0.1, \qquad \Pr(X = 2) = 0.4,$$
$$\Pr(X = 3) = 0.3, \qquad \Pr(X = 4) = 0.2.$$

Here, $\Pr(X \le 2) = 1/2$, and $\Pr(X \ge 3) = 1/2$. Therefore, every value of m in the closed interval $2 \le m \le 3$ will be a median of this distribution. The most popular choice of median of this distribution would be the midpoint 2.5. ◀

Example 4.5.3 The Median of a Continuous Distribution. Suppose that X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \begin{cases} 4x^3 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The unique median of this distribution will be the number m such that

$$\int_0^m 4x^3\,dx = \int_m^1 4x^3\,dx = \frac{1}{2}.$$

This number is $m = 1/2^{1/4}$. ◀

Example 4.5.4 A Continuous Distribution for Which the Median Is Not Unique. Suppose that X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \begin{cases} 1/2 & \text{for } 0 \le x \le 1, \\ 1 & \text{for } 2.5 \le x \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

Here, for every value of m in the closed interval $1 \le m \le 2.5$, $\Pr(X \le m) = \Pr(X \ge m) = 1/2$. Therefore, every value of m in the interval $1 \le m \le 2.5$ is a median of this distribution. ◀

Comparison of the Mean and the Median

Example 4.5.5 Last Lottery Number. In a state lottery game, a three-digit number from 000 to 999 is drawn each day. After several years, all but one of the 1000 possible numbers has been drawn. A lottery official would like to predict how much longer it will be until that missing number is finally drawn. Let X be the number of days ($X = 1$ being tomorrow) until that number appears. It is not difficult to determine the distribution of X, assuming that all 1000 numbers are equally likely to be drawn each day and

that the draws are independent. Let A_x stand for the event that the missing number is drawn on day x for $x = 1, 2, \ldots$. Then $\{X = 1\} = A_1$, and for $x > 1$,

$$\{X = x\} = A_1^c \cap \cdots \cap A_{x-1}^c \cap A_x.$$

Since the A_x events are independent and all have probability 0.001, it is easy to see that the p.f. of X is

$$f(x) = \begin{cases} 0.001(0.999)^{x-1} & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

But, the lottery official wants to give a single-number prediction for when the number will be drawn. What summary of the distribution would be appropriate for this prediction? ◀

The lottery official in Example 4.5.5 wants some sort of "average" or "middle" number to summarize the distribution of the number of days until the last number appears. Presumably she wants a prediction that is neither excessively large nor too small. Either the mean or a median of X can be used as such a summary of the distribution. Some important properties of the mean have already been described in this chapter, and several more properties will be given later in the book. However, for many purposes the median is a more useful measure of the middle of the distribution than is the mean. For example, every distribution has a median, but not every distribution has a mean. As illustrated in Example 4.3.5, the mean of a distribution can be made very large by removing a small but positive amount of probability from any part of the distribution and assigning this amount to a sufficiently large value of x. On the other hand, the median may be unaffected by a similar change in probabilities. If any amount of probability is removed from a value of x larger than the median and assigned to an arbitrarily large value of x, the median of the new distribution will be the same as that of the original distribution. In Example 4.3.5, all numbers in the interval [0, 1] are medians of both random variables X and Y despite the large difference in their means.

Example 4.5.6 Annual Incomes. Suppose that the mean annual income among the families in a certain community is \$30,000. It is possible that only a few families in the community actually have an income as large as \$30,000, but those few families have incomes that are very much larger than \$30,000. As an extreme example, suppose that there are 100 families and 99 of them have income of \$1,000 while the other one has income of \$2,901,000. If, however, the median annual income among the families is \$30,000, then at least one-half of the families must have incomes of \$30,000 or more. ◀

The median has one convenient property that the mean *does not* have.

Theorem 4.5.1 One-to-One Function. Let X be a random variable that takes values in an interval I of real numbers. Let r be a one-to-one function defined on the interval I. If m is a median of X, then $r(m)$ is a median of $r(X)$.

Proof Let $Y = r(X)$. We need to show that $\Pr(Y \geq r(m)) \geq 1/2$ and $\Pr(Y \leq r(m)) \geq 1/2$. Since r is one-to-one on the interval I, it must be either increasing or decreasing over the interval I. If r is increasing, then $Y \geq r(m)$ if and only if $X \geq m$, so $\Pr(Y \geq r(m)) = \Pr(X \geq m) \geq 1/2$. Similarly, $Y \leq r(m)$ if and only if $X \leq m$ and $\Pr(Y \leq r(m)) \geq 1/2$ also. If r is decreasing, then $Y \geq r(m)$ if and only if $X \leq m$. The remainder of the proof is then similar to the preceding. ■

We shall now consider two specific criteria by which the prediction of a random variable X can be judged. By the first criterion, the optimal prediction that can be made is the mean. By the second criterion, the optimal prediction is the median.

Minimizing the Mean Squared Error

Suppose that X is a random variable with mean μ and variance σ^2. Suppose also that the value of X is to be observed in some experiment, but this value must be predicted before the observation can be made. One basis for making the prediction is to select some number d for which the expected value of the square of the error $X - d$ will be a minimum.

Definition 4.5.2 Mean Squared Error/M.S.E.. The number $E[(X-d)^2]$ is called the *mean squared error* (M.S.E.) of the prediction d.

The next result shows that the number d for which the M.S.E. is minimized is $E(X)$.

Theorem 4.5.2 Let X be a random variable with finite variance σ^2, and let $\mu = E(X)$. For every number d,

$$E[(X-\mu)^2] \le E[(X-d)^2]. \tag{4.5.1}$$

Furthermore, there will be equality in the relation (4.5.1) if and only if $d = \mu$.

Proof For every value of d,

$$\begin{aligned} E[(X-d)^2] &= E(X^2 - 2\,dX + d^2) \\ &= E(X^2) - 2\,d\mu + d^2. \end{aligned} \tag{4.5.2}$$

The final expression in Eq. (4.5.2) is simply a quadratic function of d. By elementary differentiation it will be found that the minimum value of this function is attained when $d = \mu$. Hence, in order to minimize the M.S.E., the predicted value of X should be its mean μ. Furthermore, when this prediction is used, the M.S.E. is simply $E[(X-\mu)^2] = \sigma^2$. ■

Example 4.5.7 Last Lottery Number. In Example 4.5.5, we discussed a state lottery in which one number had never yet been drawn. Let X stand for the number of days until that last number is eventually drawn. The p.f. of X was computed in Example 4.5.5 as

$$f(x) = \begin{cases} 0.001(0.999)^{x-1} & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

We can compute the mean of X as

$$E(X) = \sum_{x=1}^{\infty} x 0.001(0.999)^{x-1} = 0.001 \sum_{x=1}^{\infty} x(0.999)^{x-1}. \tag{4.5.3}$$

At first, this sum does not look like one that is easy to compute. However, it is closely related to the general sum

$$g(y) = \sum_{x=0}^{\infty} y^x = \frac{1}{1-y},$$

if $0 < y < 1$. Using properties of power series from calculus, we know that the derivative of $g(y)$ can be found by differentiating the individual terms of the power series. That is,

$$g'(y) = \sum_{x=0}^{\infty} x y^{x-1} = \sum_{x=1}^{\infty} x y^{x-1},$$

for $0 < y < 1$. But we also know that $g'(y) = 1/(1-y)^2$. The last sum in Eq. (4.5.3) is $g'(0.999) = 1/(0.001)^2$. It follows that

$$E(X) = 0.001 \frac{1}{(0.001)^2} = 1000. \quad \blacktriangleleft$$

Minimizing the Mean Absolute Error

Another possible basis for predicting the value of a random variable X is to choose some number d for which $E(|X - d|)$ will be a minimum.

Definition 4.5.3 Mean Absolute Error/M.A.E. The number $E(|X - d|)$ is called the *mean absolute error* (M.A.E.) of the prediction d.

We shall now show that the M.A.E. is minimized when the chosen value of d is a median of the distribution of X.

Theorem 4.5.3 Let X be a random variable with finite mean, and let m be a median of the distribution of X. For every number d,

$$E(|X - m|) \le E(|X - d|). \tag{4.5.4}$$

Furthermore, there will be equality in the relation (4.5.4) if and only if d is also a median of the distribution of X.

Proof For convenience, we shall assume that X has a continuous distribution for which the p.d.f. is f. The proof for any other type of distribution is similar. Suppose first that $d > m$. Then

$$\begin{aligned} E(|X - d|) - E(|X - m|) &= \int_{-\infty}^{\infty} (|x - d| - |x - m|) f(x)\, dx \\ &= \int_{-\infty}^{m} (d - m) f(x)\, dx + \int_{m}^{d} (d + m - 2x) f(x)\, dx + \int_{d}^{\infty} (m - d) f(x)\, dx \\ &\ge \int_{-\infty}^{m} (d - m) f(x)\, dx + \int_{m}^{d} (m - d) f(x)\, dx + \int_{d}^{\infty} (m - d) f(x)\, dx \\ &= (d - m)[\Pr(X \le m) - \Pr(X > m)]. \end{aligned} \tag{4.5.5}$$

Since m is a median of the distribution of X, it follows that

$$\Pr(X \le m) \ge 1/2 \ge \Pr(X > m). \tag{4.5.6}$$

The final difference in the relation (4.5.5) is therefore nonnegative. Hence,

$$E(|X - d|) \ge E(|X - m|). \tag{4.5.7}$$

Furthermore, there can be equality in the relation (4.5.7) only if the inequalities in relations (4.5.5) and (4.5.6) are actually equalities. A careful analysis shows that these inequalities will be equalities only if d is also a median of the distribution of X.

The proof for every value of d such that $d < m$ is similar. ■

Example 4.5.8

Last Lottery Number. In Example 4.5.5, in order to compute the median of X, we must find the smallest number x such that the c.d.f. $F(x) \geq 0.5$. For integer x, we have

$$F(x) = \sum_{n=1}^{x} 0.001(0.999)^{n-1}.$$

We can use the popular formula

$$\sum_{n=0}^{x} y^n = \frac{1 - y^{x+1}}{1 - y}$$

to see that, for integer $x \geq 1$,

$$F(x) = 0.001\frac{1 - (0.999)^x}{1 - 0.999} = 1 - (0.999)^x.$$

Setting this equal to 0.5 and solving for x gives $x = 692.8$; hence, the median of X is 693. The median is unique because $F(x)$ never takes the exact value 0.5 for any integer x. The median of X is much smaller than the mean of 1000 found in Example 4.5.7. ◀

The reason that the mean is so much larger than the median in Examples 4.5.7 and 4.5.8 is that the distribution has probability at arbitrarily large values but is bounded below. The probability at these large values pulls the mean up because there is no probability at equally small values to balance. The median is not affected by how the upper half of the probability is distributed. The following example involves a symmetric distribution. Here, the mean and median(s) are more similar.

Example 4.5.9

Predicting a Discrete Uniform Random Variable. Suppose that the probability is 1/6 that a random variable X will take each of the following six values: 1, 2, 3, 4, 5, 6. We shall determine the prediction for which the M.S.E. is minimum and the prediction for which the M.A.E. is minimum.

In this example,

$$E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5.$$

Therefore, the M.S.E. will be minimized by the unique value $d = 3.5$.

Also, every number m in the closed interval $3 \leq m \leq 4$ is a median of the given distribution. Therefore, the M.A.E. will be minimized by every value of d such that $3 \leq d \leq 4$ and only by such a value of d. Because the distribution of X is symmetric, the mean of X is also a median of X. ◀

Note: When the M.A.E. and M.S.E. Are Finite. We noted that the median exists for every distribution, but the M.A.E. is finite if and only if the distribution has a finite mean. Similarly, the M.S.E. is finite if and only if the distribution has a finite variance.

Summary

A median of X is any number m such that $\Pr(X \leq m) \geq 1/2$ and $\Pr(X \geq m) \geq 1/2$. To minimize $E(|X - d|)$ by choice of d, one must choose d to be a median of X. To minimize $E[(X - d)^2]$ by choice of d, one must choose $d = E(X)$.

Exercises

1. Prove that the 1/2 quantile as defined in Definition 3.3.2 is a median as defined in Definition 4.5.1.

2. Suppose that a random variable X has a discrete distribution for which the p.f. is as follows:

$$f(x) = \begin{cases} cx & \text{for } x = 1, 2, 3, 4, 5, 6, \\ 0 & \text{otherwise.} \end{cases}$$

Determine all the medians of this distribution.

3. Suppose that a random variable X has a continuous distribution for which the p.d.f. is as follows:

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

Determine all the medians of this distribution.

4. In a small community consisting of 153 families, the number of families that have k children ($k = 0, 1, 2, \ldots$) is given in the following table:

Number of children	Number of families
0	21
1	40
2	42
3	27
4 or more	23

Determine the mean and the median of the number of children per family. (For the mean, assume that all families with four or more children have only four children. Why doesn't this point matter for the median?)

5. Suppose that an observed value of X is equally likely to come from a continuous distribution for which the p.d.f. is f or from one for which the p.d.f. is g. Suppose that $f(x) > 0$ for $0 < x < 1$ and $f(x) = 0$ otherwise, and suppose also that $g(x) > 0$ for $2 < x < 4$ and $g(x) = 0$ otherwise. Determine: **(a)** the mean and
(b) the median of the distribution of X.

6. Suppose that a random variable X has a continuous distribution for which the p.d.f. f is as follows:

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of d that minimizes
(a) $E[(X - d)^2]$ and **(b)** $E(|X - d|)$.

7. Suppose that a person's score X on a certain examination will be a number in the interval $0 \le X \le 1$ and that X has a continuous distribution for which the p.d.f. is as follows:

$$f(x) = \begin{cases} x + \frac{1}{2} & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the prediction of X that minimizes **(a)** the M.S.E. and **(b)** the M.A.E.

8. Suppose that the distribution of a random variable X is symmetric with respect to the point $x = 0$ and that $E(X^4) < \infty$. Show that $E[(X - d)^4]$ is minimized by the value $d = 0$.

9. Suppose that a fire can occur at any one of five points along a road. These points are located at -3, -1, 0, 1, and 2 in Fig. 4.9. Suppose also that the probability that each of these points will be the location of the next fire that occurs along the road is as specified in Fig. 4.9.

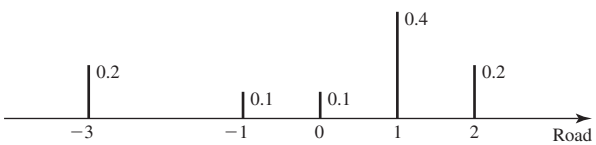

Figure 4.9 Probabilities for Exercise 9.

a. At what point along the road should a fire engine wait in order to minimize the expected value of the square of the distance that it must travel to the next fire?

b. Where should the fire engine wait to minimize the expected value of the distance that it must travel to the next fire?

10. If n houses are located at various points along a straight road, at what point along the road should a store be located in order to minimize the sum of the distances from the n houses to the store?

11. Let X be a random variable having the binomial distribution with parameters $n = 7$ and $p = 1/4$, and let Y be a random variable having the binomial distribution with parameters $n = 5$ and $p = 1/2$. Which of these two random variables can be predicted with the smaller M.S.E.?

12. Consider a coin for which the probability of obtaining a head on each given toss is 0.3. Suppose that the coin is to be tossed 15 times, and let X denote the number of heads that will be obtained.

a. What prediction of X has the smallest M.S.E.?

b. What prediction of X has the smallest M.A.E.?

13. Suppose that the distribution of X is symmetric around a point m. Prove that m is a median of X.

14. Find the median of the Cauchy distribution defined in Example 4.1.8.

15. Let X be a random variable with c.d.f. F. Suppose that $a < b$ are numbers such that both a and b are medians of X.

a. Prove that $F(a) = 1/2$.

b. Prove that there exist a smallest $c \le a$ and a largest $d \ge b$ such that every number in the closed interval $[c, d]$ is a median of X.

c. If X has a discrete distribution, prove that $F(d) > 1/2$.

16. Let X be a random variable. Suppose that there exists a number m such that $\Pr(X < m) = \Pr(X > m)$. Prove that m is a median of the distribution of X.

17. Let X be a random variable. Suppose that there exists a number m such that $\Pr(X < m) < 1/2$ and $\Pr(X > m) < 1/2$. Prove that m is the unique median of the distribution of X.

18. Prove the following extension of Theorem 4.5.1. Let m be the p quantile of the random variable X. (See Definition 3.3.2.) If r is a strictly increasing function, then $r(m)$ is the p quantile of $r(X)$.

4.6 Covariance and Correlation

When we are interested in the joint distribution of two random variables, it is useful to have a summary of how much the two random variables depend on each other. The covariance and correlation are attempts to measure that dependence, but they only capture a particular type of dependence, namely linear dependence.

Covariance

Example 4.6.1

Test Scores. When applying for college, high school students often take a number of standardized tests. Consider a particular student who will take both a verbal and a quantitative test. Let X be this student's score on the verbal test, and let Y be the same student's score on the quantitative test. Although there are students who do much better on one test than the other, it might still be reasonable to expect that a student who does very well on one test to do at least a little better than average on the other. We would like to find a numerical summary of the joint distribution of X and Y that reflects the degree to which we believe a high or low score on one test will be accompanied by a high or low score on the other test. ◀

When we consider the joint distribution of two random variables, the means, the medians, and the variances of the variables provide useful information about their marginal distributions. However, these values do not provide any information about the relationship between the two variables or about their tendency to vary together rather than independently. In this section and the next one, we shall introduce summaries of a joint distribution that enable us to measure the association between two random variables, determine the variance of the sum of an arbitrary number of dependent random variables, and predict the value of one random variable by using the observed value of some other related variable.

Definition 4.6.1

Covariance. Let X and Y be random variables having finite means. Let $E(X) = \mu_X$ and $E(Y) = \mu_Y$ The *covariance of X and Y*, which is denoted by $\mathrm{Cov}(X, Y)$, is defined as

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)], \tag{4.6.1}$$

if the expectation in Eq. (4.6.1) exists.

It can be shown (see Exercise 2 at the end of this section) that if both X and Y have finite variance, then the expectation in Eq. (4.6.1) will exist and $\text{Cov}(X, Y)$ will be finite. However, the value of $\text{Cov}(X, Y)$ can be positive, negative, or zero.

Example 4.6.2 Test Scores. Let X and Y be the test scores in Example 4.6.1, and suppose that they have the joint p.d.f.

$$f(x, y) = \begin{cases} 2xy + 0.5 & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

We shall compute the covariance $\text{Cov}(X, Y)$. First, we shall compute the means μ_X and μ_Y of X and Y, respectively. The symmetry in the joint p.d.f. means that X and Y have the same marginal distribution; hence, $\mu_X = \mu_Y$. We see that

$$\begin{aligned} \mu_X &= \int_0^1 \int_0^1 [2x^2 y + 0.5x] dy dx \\ &= \int_0^1 [x^2 + 0.5x] dx = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}, \end{aligned}$$

so that $\mu_Y = 7/12$ as well. The covariance can be computed using Theorem 4.1.2. Specifically, we must evaluate the integral

$$\int_0^1 \int_0^1 \left(x - \frac{7}{12}\right)\left(y - \frac{7}{12}\right)(2xy + 0.5)\, dy\, dx.$$

This integral is straightforward, albeit tedious, to compute, and the result is $\text{Cov}(X, Y) = 1/144$. ◀

The following result often simplifies the calculation of a covariance.

Theorem 4.6.1 For all random variables X and Y such that $\sigma_X^2 < \infty$ and $\sigma_Y^2 < \infty$,

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y). \tag{4.6.2}$$

Proof It follows from Eq. (4.6.1) that

$$\begin{aligned} \text{Cov}(X, Y) &= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y. \end{aligned}$$

Since $E(X) = \mu_X$ and $E(Y) = \mu_Y$, Eq. (4.6.2) is obtained. ■

The covariance between X and Y is intended to measure the degree to which X and Y tend to be large at the same time or the degree to which one tends to be large while the other is small. Some intution about this interpretation can be gathered from a careful look at Eq. (4.6.1). For example, suppose that $\text{Cov}(X, Y)$ is positive. Then $X > \mu_X$ and $Y > \mu_Y$ must occur together and/or $X < \mu_X$ and $Y < \mu_Y$ must occur together to a larger extent than $X < \mu_X$ occurs with $Y > \mu_Y$ and $X > \mu_X$ occurs with $Y < \mu_Y$. Otherwise, the mean would be negative. Similarly, if $\text{Cov}(X, Y)$ is negative, then $X > \mu_X$ and $Y < \mu_Y$ must occur together and/or $X < \mu_X$ and $Y > \mu_Y$ must occur together to larger extent than the other two inequalities. If $\text{Cov}(X, Y) = 0$, then the extent to which X and Y are on the same sides of their respective means exactly balances the extent to which they are on opposite sides of their means.

Correlation

Although $\text{Cov}(X, Y)$ gives a numerical measure of the degree to which X and Y vary together, the magnitude of $\text{Cov}(X, Y)$ is also influenced by the overall magnitudes of X and Y. For example, in Exercise 5 in this section, you can prove that $\text{Cov}(2X, Y) = 2\,\text{Cov}(X, Y)$. In order to obtain a measure of association between X and Y that is *not driven by arbitrary changes in the scales* of one or the other random variable, we define a slightly different quantity next.

Definition 4.6.2 Correlation. Let X and Y be random variables with finite variances σ_X^2 and σ_Y^2, respectively. Then the *correlation of X and Y*, which is denoted by $\rho(X, Y)$, is defined as follows:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{4.6.3}$$

In order to determine the range of possible values of the correlation $\rho(X, Y)$, we shall need the following result.

Theorem 4.6.2 Schwarz Inequality. For all random variables U and V such that $E(UV)$ exists,

$$[E(UV)]^2 \le E(U^2)E(V^2). \tag{4.6.4}$$

If, in addition, the right-hand side of Eq. (4.6.4) is finite, then the two sides of Eq. (4.6.4) equal the same value if and only if there are nonzero constants a and b such that $aU + bV = 0$ with probability 1.

Proof If $E(U^2) = 0$, then $\Pr(U = 0) = 1$. Therefore, it must also be true that $\Pr(UV = 0) = 1$. Hence, $E(UV) = 0$, and the relation (4.6.4) is satisfied. Similarly, if $E(V^2) = 0$, then the relation (4.6.4) will be satisfied. Moreover, if either $E(U^2)$ or $E(V^2)$ is infinite, then the right side of the relation (4.6.4) will be infinite. In this case, the relation (4.6.4) will surely be satisfied.

For the rest of the proof, assume that $0 < E(U^2) < \infty$ and $0 < E(V^2) < \infty$. For all numbers a and b,

$$0 \le E[(aU + bV)^2] = a^2E(U^2) + b^2E(V^2) + 2abE(UV) \tag{4.6.5}$$

and

$$0 \le E[(aU - bV)^2] = a^2E(U^2) + b^2E(V^2) - 2abE(UV). \tag{4.6.6}$$

If we let $a = [E(V^2)]^{1/2}$ and $b = [E(U^2)]^{1/2}$, then it follows from the relation (4.6.5) that

$$E(UV) \ge -[E(U^2)E(V^2)]^{1/2}. \tag{4.6.7}$$

It also follows from the relation (4.6.6) that

$$E(UV) \le [E(U^2)E(V^2)]^{1/2}. \tag{4.6.8}$$

These two relations together imply that the relation (4.6.4) is satisfied.

Finally, suppose that the right-hand side of Eq. (4.6.4) is finite. Both sides of (4.6.4) equal the same value if and only if the same is true for either (4.6.7) or (4.6.8). Both sides of (4.6.7) equal the same value if and only if the rightmost expression in (4.6.5) is 0. This, in turn, is true if and only if $E[(aU + bV)^2] = 0$, which occurs if and only if $aU + bV = 0$ with probability 1. The reader can easily check that both sides of (4.6.8) equal the same value if and only if $aU - bV = 0$ with probability 1. ■

A slight variant on Theorem 4.6.2 is the result we want.

Theorem 4.6.3 Cauchy-Schwarz Inequality. Let X and Y be random variables with finite variance. Then

$$[\text{Cov}(X, Y)]^2 \le \sigma_X^2 \sigma_Y^2, \tag{4.6.9}$$

and

$$-1 \le \rho(X, Y) \le 1. \tag{4.6.10}$$

Furthermore, the inequality in Eq. (4.6.9) is an equality if and only if there are nonzero constants a and b and a constant c such that $aX + bY = c$ with probability 1.

Proof Let $U = X - \mu_X$ and $V = Y - \mu_Y$. Eq. (4.6.9) now follows directly from Theorem 4.6.2. In turn, it follows from Eq. (4.6.3) that $[\rho(X, Y)]^2 \le 1$ or, equivalently, that Eq. (4.6.10) holds. The final claim follows easily from the similar claim at the end of Theorem 4.6.2. ■

Definition 4.6.3 Positively/Negatively Correlated/Uncorrelated. It is said that X and Y are *positively correlated* if $\rho(X, Y) > 0$, that X and Y are *negatively correlated* if $\rho(X, Y) < 0$, and that X and Y are *uncorrelated* if $\rho(X, Y) = 0$.

It can be seen from Eq. (4.6.3) that $\text{Cov}(X, Y)$ and $\rho(X, Y)$ must have the same sign; that is, both are positive, or both are negative, or both are zero.

Example 4.6.3 Test Scores. For the two test scores in Example 4.6.2, we can compute the correlation $\rho(X, Y)$. The variances of X and Y are both equal to 11/144, so the correlation is $\rho(X, Y) = 1/11$. ◀

Properties of Covariance and Correlation

We shall now present four theorems pertaining to the basic properties of covariance and correlation.

The first theorem shows that independent random variables must be uncorrelated.

Theorem 4.6.4 If X and Y are independent random variables with $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$, then

$$\text{Cov}(X, Y) = \rho(X, Y) = 0.$$

Proof If X and Y are independent, then $E(XY) = E(X)E(Y)$. Therefore, by Eq. (4.6.2), $\text{Cov}(X, Y) = 0$. Also, it follows that $\rho(X, Y) = 0$. ■

The converse of Theorem 4.6.4 is not true as a general rule. Two dependent random variables can be uncorrelated. Indeed, even though Y is an explicit function of X, it is possible that $\rho(X, Y) = 0$, as in the following examples.

Example 4.6.4 Dependent but Uncorrelated Random Variables. Suppose that the random variable X can take only the three values -1, 0, and 1, and that each of these three values has the same probability. Also, let the random variable Y be defined by the relation $Y = X^2$. We shall show that X and Y are dependent but uncorrelated.

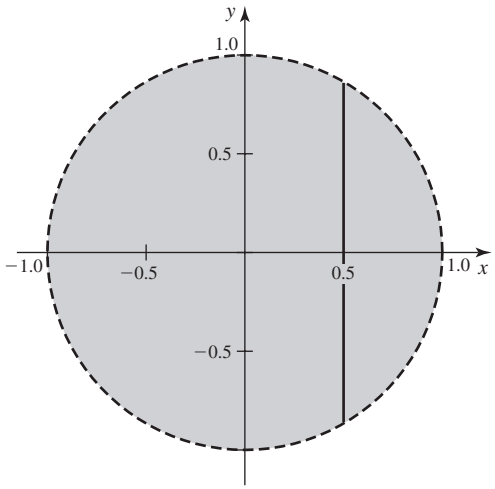

Figure 4.10 The shaded region is where the joint p.d.f. of (X, Y) is constant and nonzero in Example 4.6.5. The vertical line indicates the values of Y that are possible when $X = 0.5$.

In this example, X and Y are clearly dependent, since Y is not constant and the value of Y is completely determined by the value of X. However,

$$E(XY) = E(X^3) = E(X) = 0,$$

because X^3 is the same random variable as X. Since $E(XY) = 0$ and $E(X)E(Y) = 0$, it follows from Theorem 4.6.1 that $\text{Cov}(X, Y) = 0$ and that X and Y are uncorrelated. ◀

Example 4.6.5 Uniform Distribution Inside a Circle. Let (X, Y) have joint p.d.f. that is constant on the interior of the unit circle, the shaded region in Fig. 4.10. The constant value of the p.d.f. is one over the area of the circle, that is, $1/(2\pi)$. It is clear that X and Y are dependent since the region where the joint p.d.f. is nonzero is not a rectangle. In particular, notice that the set of possible values for Y is the interval $(-1, 1)$, but when $X = 0.5$, the set of possible values for Y is the smaller interval $(-0.866, 0.866)$. The symmetry of the circle makes it clear that both X and Y have mean 0. Also, it is not difficult to see that $E(XY) = \int\int xyf(x, y)dxdy = 0$. To see this, notice that the integral of xy over the top half of the circle is exactly the negative of the integral of xy over the bottom half. Hence, $\text{Cov}(X, Y) = 0$, but the random variables are dependent. ◀

The next result shows that if Y is a *linear* function of X, then X and Y must be correlated and, in fact, $|\rho(X, Y)| = 1$.

Theorem 4.6.5 Suppose that X is a random variable such that $0 < \sigma_X^2 < \infty$, and $Y = aX + b$ for some constants a and b, where $a \neq 0$. If $a > 0$, then $\rho(X, Y) = 1$. If $a < 0$, then $\rho(X, Y) = -1$.

Proof If $Y = aX + b$, then $\mu_Y = a\mu_X + b$ and $Y - \mu_Y = a(X - \mu_X)$. Therefore, by Eq. (4.6.1),

$$\text{Cov}(X, Y) = aE[(X - \mu_X)^2] = a\sigma_X^2.$$

Since $\sigma_Y = |a|\sigma_X$, the theorem follows from Eq. (4.6.3). ■

There is a converse to Theorem 4.6.5. That is, $|\rho(X, Y)| = 1$ implies that X and Y are linearly related. (See Exercise 17.) In general, the value of $\rho(X, Y)$ provides a measure of the extent to which two random variables X and Y are linearly related. If

the joint distribution of X and Y is relatively concentrated around a straight line in the xy-plane that has a positive slope, then $\rho(X, Y)$ will typically be close to 1. If the joint distribution is relatively concentrated around a straight line that has a negative slope, then $\rho(X, Y)$ will typically be close to -1. We shall not discuss these concepts further here, but we shall consider them again when the bivariate normal distribution is introduced and studied in Sec. 5.10.

Note: Correlation Measures Only Linear Relationship. A large value of $|\rho(X, Y)|$ means that X and Y are close to being linearly related and hence are closely related. But a small value of $|\rho(X, Y)|$ does not mean that X and Y are not close to being related. Indeed, Example 4.6.4 illustrates random variables that are functionally related but have 0 correlation.

We shall now determine the variance of the sum of random variables that are not necessarily independent.

Theorem 4.6.6 If X and Y are random variables such that $\text{Var}(X) < \infty$ and $\text{Var}(Y) < \infty$, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y). \tag{4.6.11}$$

Proof Since $E(X + Y) = \mu_X + \mu_Y$, then

$$\begin{aligned}\text{Var}(X + Y) &= E[(X + Y - \mu_X - \mu_Y)^2] \\ &= E[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \\ &= \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y).\end{aligned}$$

■

For all constants a and b, it can be shown that $\text{Cov}(aX, bY) = ab\,\text{Cov}(X, Y)$ (see Exercise 5 at the end of this section). The following then follows easily from Theorem 4.6.6.

Corollary 4.6.1 Let a, b, and c be constants. Under the conditions of Theorem 4.6.6,

$$\text{Var}(aX + bY + c) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y). \tag{4.6.12}$$

■

A particularly useful special case of Corollary 4.6.1 is

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y). \tag{4.6.13}$$

Example 4.6.6 Investment Portfolio. Consider, once again, the investor in Example 4.3.7 on page 230 trying to choose a portfolio with \$100,000 to invest. We shall make the same assumptions about the returns on the two stocks, except that now we will suppose that the correlation between the two returns R_1 and R_2 is -0.3, reflecting a belief that the two stocks tend to react in opposite ways to common market forces. The variance of a portfolio of s_1 shares of the first stock, s_2 shares of the second stock, and s_3 dollars invested at 3.6% is now

$$\text{Var}(s_1R_1 + s_2R_2 + 0.036s_3) = 55s_1^2 + 28s_2^2 - 0.3\sqrt{55 \times 28}s_1s_2.$$

We continue to assume that (4.3.2) holds. Figure 4.11 shows the relationship between the mean and variance of the efficient portfolios in this example and Example 4.3.7. Notice how the variances are smaller in this example than in Example 4.3.7. This is due to the fact that the negative correlation lowers the variance of a linear combination with positive coefficients. ◀

Theorem 4.6.6 can also be extended easily to the variance of the sum of n random variables, as follows.

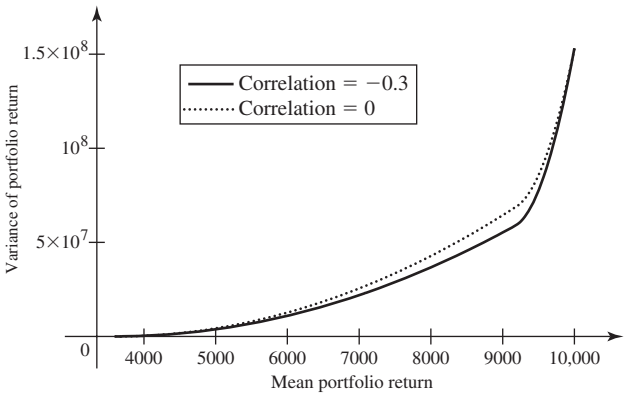

Figure 4.11 Mean and variance of efficient investment portfolios.

Theorem 4.6.7 If $X_1, \ldots, X_n$ are random variables such that $\text{Var}(X_i) < \infty$ for $i = 1, \ldots, n$, then

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum\sum_{i<j} \text{Cov}(X_i, X_j). \tag{4.6.14}$$

Proof For every random variable Y, $\text{Cov}(Y, Y) = \text{Var}(Y)$. Therefore, by using the result in Exercise 8 at the end of this section, we can obtain the following relation:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j).$$

We shall separate the final sum in this relation into two sums: (i) the sum of those terms for which $i = j$ and (ii) the sum of those terms for which $i \neq j$. Then, if we use the fact that $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, we obtain the relation

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

■

The following is a simple corrolary to Theorem 4.6.7.

Corollary 4.6.2 If $X_1, \ldots, X_n$ are uncorrelated random variables (that is, if X_i and X_j are uncorrelated whenever $i \neq j$), then

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \tag{4.6.15}$$

■

Corollary 4.6.2 extends Theorem 4.3.5 on page 230, which states that (4.6.15) holds if $X_1, \ldots, X_n$ are independent random variables.

Note: In General, Variances Add Only for Uncorrelated Random Variables. The variance of a sum of random variables should be calculated using Theorem 4.6.7 in general. Corollary 4.6.2 applies only for uncorrelated random variables.

Summary

The covariance of X and Y is $\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$. The correlation is $\rho(X, Y) = \text{Cov}(X, Y)/[\text{Var}(X)\,\text{Var}(Y)]^{1/2}$, and it measures the extent to which X and Y are linearly related. Indeed, X and Y are precisely linearly related if and only if $|\rho(X, Y)| = 1$. The variance of a sum of random variables can be expressed as the sum of the variances plus two times the sum of the covariances. The variance of a linear function is $\text{Var}(aX + bY + c) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y)$.

Exercises

1. Suppose that the pair (X, Y) is uniformly distributed on the interior of a circle of radius 1. Compute $\rho(X, Y)$.

2. Prove that if $\text{Var}(X) < \infty$ and $\text{Var}(Y) < \infty$, then $\text{Cov}(X, Y)$ is finite. *Hint:* By considering the relation $[(X - \mu_X) \pm (Y - \mu_Y)]^2 \geq 0$, show that

$$|(X - \mu_X)(Y - \mu_Y)| \leq \frac{1}{2}[(X - \mu_X)^2 + (Y - \mu_Y)^2].$$

3. Suppose that X has the uniform distribution on the interval $[-2, 2]$ and $Y = X^6$. Show that X and Y are uncorrelated.

4. Suppose that the distribution of a random variable X is symmetric with respect to the point $x = 0, 0 < E(X^4) < \infty$, and $Y = X^2$. Show that X and Y are uncorrelated.

5. For all random variables X and Y and all constants a, b, c, and d, show that

$$\text{Cov}(aX + b, cY + d) = ac\,\text{Cov}(X, Y).$$

6. Let X and Y be random variables such that $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$. Suppose that $U = aX + b$ and $V = cY + d$, where $a \neq 0$ and $c \neq 0$. Show that $\rho(U, V) = \rho(X, Y)$ if $ac > 0$, and $\rho(U, V) = -\rho(X, Y)$ if $ac < 0$.

7. Let X, Y, and Z be three random variables such that $\text{Cov}(X, Z)$ and $\text{Cov}(Y, Z)$ exist, and let a, b, and c be arbitrary given constants. Show that

$$\text{Cov}(aX + bY + c, Z) = a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z).$$

8. Suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are random variables such that $\text{Cov}(X_i, Y_j)$ exists for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, and suppose that $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ are constants. Show that

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

9. Suppose that X and Y are two random variables, which may be dependent, and $\text{Var}(X) = \text{Var}(Y)$. Assuming that $0 < \text{Var}(X + Y) < \infty$ and $0 < \text{Var}(X - Y) < \infty$, show that the random variables $X + Y$ and $X - Y$ are uncorrelated.

10. Suppose that X and Y are negatively correlated. Is $\text{Var}(X + Y)$ larger or smaller than $\text{Var}(X - Y)$?

11. Show that two random variables X and Y cannot possibly have the following properties: $E(X) = 3$, $E(Y) = 2$, $E(X^2) = 10$, $E(Y^2) = 29$, and $E(XY) = 0$.

12. Suppose that X and Y have a continuous joint distribution for which the joint p.d.f. is as follows:

$$f(x, y) = \begin{cases} \frac{1}{3}(x + y) & \text{for } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the value of $\text{Var}(2X - 3Y + 8)$.

13. Suppose that X and Y are random variables such that $\text{Var}(X) = 9$, $\text{Var}(Y) = 4$, and $\rho(X, Y) = -1/6$. Determine **(a)** $\text{Var}(X + Y)$ and **(b)** $\text{Var}(X - 3Y + 4)$.

14. Suppose that X, Y, and Z are three random variables such that $\text{Var}(X) = 1$, $\text{Var}(Y) = 4$, $\text{Var}(Z) = 8$, $\text{Cov}(X, Y) = 1$, $\text{Cov}(X, Z) = -1$, and $\text{Cov}(Y, Z) = 2$. Determine **(a)** $\text{Var}(X + Y + Z)$ and **(b)** $\text{Var}(3X - Y - 2Z + 1)$.

15. Suppose that $X_1, \ldots, X_n$ are random variables such that the variance of each variable is 1 and the correlation between each pair of different variables is $1/4$. Determine $\text{Var}(X_1 + \cdots + X_n)$.

16. Consider the investor in Example 4.2.3 on page 220. Suppose that the returns R_1 and R_2 on the two stocks have correlation -1. A portfolio will consist of s_1 shares of the first stock and s_2 shares of the second stock where $s_1, s_2 \geq 0$. Find a portfolio such that the total cost of the portfolio is \$6000 and the variance of the return is 0. Why is this situation unrealistic?

17. Let X and Y be random variables with finite variance. Prove that $|\rho(X, Y)| = 1$ implies that there exist constants a, b, and c such that $aX + bY = c$ with probability 1. *Hint:* Use Theorem 4.6.2 with $U = X - \mu_X$ and $V = Y - \mu_Y$.

18. Let X and Y have a continuous distribution with joint p.d.f.

$$f(x, y) = \begin{cases} x + y & \text{for } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the covariance $\text{Cov}(X, Y)$.

4.7 Conditional Expectation

Since expectations (including variances and covariances) are properties of distributions, there will exist conditional versions of all such distributional summaries as well as conditional versions of all theorems that we have proven or will later prove about expectations. In particular, suppose that we wish to predict one random variable Y using a function $d(X)$ of another random variable X so as to minimize $E([Y - d(X)]^2)$. Then $d(X)$ should be the conditional mean of Y given X. There is also a very useful theorem that is an extension to expectations of the law of total probability.

Definition and Basic Properties

Example 4.7.1

Household Survey. A collection of households were surveyed, and each household reported the number of members and the number of automobiles owned. The reported numbers are in Table 4.1.

Suppose that we were to sample a household at random from those households in the survey and learn the number of members. What would then be the expected number of automobiles that they own? ◀

The question at the end of Example 4.7.1 is closely related to the conditional distribution of one random variable given the other, as defined in Sec. 3.6.

Definition 4.7.1

Conditional Expectation/Mean. Let X and Y be random variables such that the mean of Y exists and is finite. The *conditional expectation (or conditional mean) of Y given $X = x$* is denoted by $E(Y|x)$ and is defined to be the expectation of the conditional distribution of Y given $X = x$.

For example, if Y has a continuous conditional distribution given $X = x$ with conditional p.d.f. $g_2(y|x)$, then

$$E(Y|x) = \int_{-\infty}^{\infty} y g_2(y|x)\, dy. \tag{4.7.1}$$

Similarly, if Y has a discrete conditional distribution given $X = x$ with conditional p.f. $g_2(y|x)$, then

$$E(Y|x) = \sum_{\text{All } y} y g_2(y|x). \tag{4.7.2}$$

Table 4.1 Reported numbers of household members and automobiles in Example 4.7.1

Number of automobiles	Number of members							
	1	2	3	4	5	6	7	8
0	10	7	3	2	2	1	0	0
1	12	21	25	30	25	15	5	1
2	1	5	10	15	20	11	5	3
3	0	2	3	5	5	3	2	1

The value of $E(Y|x)$ will not be uniquely defined for those values of x such that the marginal p.f. or p.d.f. of X satisfies $f_1(x) = 0$. However, since these values of x form a set of points whose probability is 0, the definition of $E(Y|x)$ at such a point is irrelevant. (See Exercise 11 in Sec. 3.6.) It is also possible that there will be some values of x such that the mean of the conditional distribution of Y given $X = x$ is undefined for those x values. When the mean of Y exists and is finite, the set of x values for which the conditional mean is undefined has probability 0.

The expressions in Eqs. (4.7.1) and (4.7.2) are functions of x. These functions of x can be computed before X is observed, and this idea leads to the following useful concept.

Definition 4.7.2 Conditional Means as Random Variables. Let $h(x)$ stand for the function of x that is denoted $E(Y|x)$ in either (4.7.1) or (4.7.2). Define the symbol $E(Y|X)$ to mean $h(X)$ and call it the *conditional mean of Y given X*.

In other words, $E(Y|X)$ is a random variable (a function of X) whose value when $X = x$ is $E(Y|x)$. Obviously, we could define $E(X|Y)$ and $E(X|y)$ analogously.

Example 4.7.2 Household Survey. Consider the household survey in Example 4.7.1. Let X be the number of members in a randomly selected household from the survey, and let Y be the number of cars owned by that household. The 250 surveyed households are all equally likely to be selected, so $\Pr(X = x, Y = y)$ is the number of households with x members and y cars, divided by 250. Those probabilities are reported in Table 4.2. Suppose that the sampled household has $X = 4$ members. The conditional p.f. of Y given $X = 4$ is $g_2(y|4) = f(4, y)/f_1(4)$, which is the $x = 4$ column of Table 4.2 divided by $f_1(4) = 0.208$, namely,

$$g_2(0|4) = 0.0385, \quad g_2(1|4) = 0.5769, \quad g_2(2|4) = 0.2885, \quad g_2(3|4) = 0.0962.$$

The conditional mean of Y given $X = 4$ is then

$$E(Y|4) = 0 \times 0.0385 + 1 \times 0.5769 + 2 \times 0.2885 + 3 \times 0.0962 = 1.442.$$

Similarly, we can compute $E(Y|x)$ for all eight values of x. They are

x	1	2	3	4	5	6	7	8
$E(Y\|x)$	0.609	1.057	1.317	1.442	1.538	1.533	1.75	2

Table 4.2 Joint p.f. $f(x, y)$ of X and Y in Example 4.7.2 together with marginal p.f.'s $f_1(x)$ and $f_2(y)$

	x								
y	1	2	3	4	5	6	7	8	$f_2(y)$
0	0.040	0.028	0.012	0.008	0.008	0.004	0	0	0.100
1	0.048	0.084	0.100	0.120	0.100	0.060	0.020	0.004	0.536
2	0.004	0.020	0.040	0.060	0.080	0.044	0.020	0.012	0.280
3	0	0.008	0.012	0.020	0.020	0.012	0.008	0.004	0.084
$f_1(x)$	0.092	0.140	0.164	0.208	0.208	0.120	0.048	0.020	

The random variable that takes the value 0.609 when the sampled household has one member, takes the value 1.057 when the sampled household has two members, and so on, is the random variable $E(Y|X)$. ◀

Example 4.7.3

A Clinical Trial. Consider a clinical trial in which a number of patients will be treated and each patient will have one of two possible outcomes: success or failure. Let P be the proportion of successes in a very large collection of patients, and let $X_i = 1$ if the ith patient is a success and $X_i = 0$ if not. Assume that the random variables $X_1, X_2, \ldots$ are conditionally independent given $P = p$ with $\Pr(X_i = 1|P = p) = p$. Let $X = X_1 + \cdots + X_n$, which is the number of patients out of the first n who are successes. We now compute the conditional mean of X given P. The patients are independent and identically distributed conditional on $P = p$. Hence, the conditional distribution of X given $P = p$ is the binomial distribution with parameters n and p. As we saw in Sec. 4.2, the mean of this binomial distribution is np, so $E(X|p) = np$ and $E(X|P) = nP$. Later, we will show how to compute the conditional mean of P given X. This can be used to predict P after observing X. ◀

Note: The Conditional Mean of Y Given X Is a Random Variable. Because $E(Y|X)$ is a function of the random variable X, it is itself a random variable with its own probability distribution, which can be derived from the distribution of X. On the other hand, $h(x) = E(Y|x)$ is a function of x that can be manipulated like any other function. The connection between the two is that when one substitutes the random variable X for x in $h(x)$, the result is $h(X) = E(Y|X)$.

We shall now show that the mean of the random variable $E(Y|X)$ must be $E(Y)$. A similar calculation shows that the mean of $E(X|Y)$ must be $E(X)$.

Theorem 4.7.1

Law of Total Probability for Expectations. Let X and Y be random variables such that Y has finite mean. Then

$$E[E(Y|X)] = E(Y). \tag{4.7.3}$$

Proof We shall assume, for convenience, that X and Y have a continuous joint distribution. Then

$$\begin{aligned} E[E(Y|X)] &= \int_{-\infty}^{\infty} E(Y|x) f_1(x)\, dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y g_2(y|x) f_1(x)\, dy\, dx. \end{aligned}$$

Since $g_2(y|x) = f(x, y)/f_1(x)$, it follows that

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

The proof for a discrete distribution or a more general type of distribution is similar. ■

Example 4.7.4

Household Survey. At the end of Example 4.7.2, we described the random variable $E(Y|X)$. Its distribution can be constructed from that description. It has a discrete distribution that takes the eight values of $E(Y|x)$ listed near the end of that example with corresponding probabilities $f_1(x)$ for $x = 1, \ldots, 8$. To be specific, let $Z = E(Y|X)$, then $\Pr[Z = E(Y|x)] = f_1(x)$ for $x = 1, \ldots, 8$. The specific values are

z	0.609	1.057	1.317	1.442	1.538	1.533	1.75	2
$\Pr(Z = z)$	0.092	0.140	0.164	0.208	0.208	0.120	0.048	0.020

We can compute $E(Z) = 0.609 \times 0.092 + \cdots + 2 \times 0.020 = 1.348$. The reader can verify that $E(Y) = 1.348$ by using the values of $f_2(y)$ in Table 4.2. ◀

Example 4.7.5

A Clinical Trial. In Example 4.7.3, we let X be the number of patients out of the first n who are successes. The conditional mean of X given $P = p$ was computed as $E(X|p) = np$, where P is the proportion of successes in a large population of patients. If the distribution of P is uniform on the interval [0, 1], then the marginal expected value of X is $E[E(X|P)] = E(nP) = n/2$. We will see how to calculate $E(P|X)$ in Example 4.7.8. ◀

Example 4.7.6

Choosing Points from Uniform Distributions. Suppose that a point X is chosen in accordance with the uniform distribution on the interval [0, 1]. Also, suppose that after the value $X = x$ has been observed ($0 < x < 1$), a point Y is chosen in accordance with a uniform distribution on the interval $[x, 1]$. We shall determine the value of $E(Y)$.

For each given value of x $(0 < x < 1)$, $E(Y|x)$ will be equal to the midpoint $(1/2)(x + 1)$ of the interval $[x, 1]$. Therefore, $E(Y|X) = (1/2)(X + 1)$ and

$$E(Y) = E[E(Y|X)] = \frac{1}{2}[E(X) + 1] = \frac{1}{2}\left(\frac{1}{2} + 1\right) = \frac{3}{4}.$$ ◀

When manipulating the conditional distribution given $X = x$, it is safe to act as if X is the constant x. This fact, which can simplify the calculation of certain conditional means, is now stated without proof.

Theorem 4.7.2

Let X and Y be random variables, and let $Z = r(X, Y)$ for some function r. The conditional distribution of Z given $X = x$ is the same as the conditional distribution of $r(x, Y)$ given $X = x$. ■

One consequence of Theorem 4.7.2 when X and Y have a continuous joint distribution is that

$$E(Z|x) = E(r(x, Y)|x) = \int_{-\infty}^{\infty} r(x, y) g_2(y|x)\, dy.$$

Theorem 4.7.1 also implies that for two arbitrary random variables X and Y,

$$E\{E[r(X, Y)|X]\} = E[r(X, Y)], \tag{4.7.4}$$

by letting $Z = r(X, Y)$ and noting that $E\{E(Z|X)\} = E(Z)$.

We can define, in a similar manner, the conditional expectation of $r(X, Y)$ given Y and the conditional expectation of a function $r(X_1, \ldots, X_n)$ of several random variables given one or more of the variables $X_1, \ldots, X_n$.

Example 4.7.7

Linear Conditional Expectation. Suppose that $E(Y|X) = aX + b$ for some constants a and b. We shall determine the value of $E(XY)$ in terms of $E(X)$ and $E(X^2)$.

By Eq. (4.7.4), $E(XY) = E[E(XY|X)]$. Furthermore, since X is considered to be given and fixed in the conditional expectation,

$$E(XY|X) = XE(Y|X) = X(aX + b) = aX^2 + bX.$$

Therefore,

$$E(XY) = E(aX^2 + bX) = aE(X^2) + bE(X). \quad \blacktriangleleft$$

The mean is not the only feature of a conditional distribution that is important enough to get its own name.

Definition 4.7.3 Conditional Variance. For every given value x, let $\mathrm{Var}(Y|x)$ denote the variance of the conditional distribution of Y given that $X = x$. That is,

$$\mathrm{Var}(Y|x) = E\{[Y - E(Y|x)]^2|x\}. \tag{4.7.5}$$

We call $\mathrm{Var}(Y|x)$ the *conditional variance of Y given $X = x$*.

The expression in Eq. (4.7.5) is once again a function $v(x)$. We shall define $\mathrm{Var}(Y|X)$ to be $v(X)$ and call it the *conditional variance of Y given X*.

Note: Other Conditional Quantities. In much the same way as in Definitions 4.7.1 and 4.7.3, we could define any conditional summary of a distribution that we wish. For example, conditional quantiles of Y given $X = x$ are the quantiles of the conditional distribution of Y given $X = x$. The conditional m.g.f. of Y given $X = x$ is the m.g.f. of the conditional distribution of Y given $X = x$, etc.

Prediction

At the end of Example 4.7.3, we considered the problem of predicting the proportion P of successes in a large population of patients given the observed number X of succeses in a sample of size n. In general, consider two arbitrary random variables X and Y that have a specified joint distribution and suppose that after the value of X has been observed, the value of Y must be predicted. In other words, the predicted value of Y can depend on the value of X. We shall assume that this predicted value $d(X)$ must be chosen so as to minimize the mean squared error $E\{[Y - d(X)]^2\}$.

Theorem 4.7.3 The prediction $d(X)$ that minimizes $E\{[Y - d(X)]^2\}$ is $d(X) = E(Y|X)$.

Proof We shall prove the theorem in the case in which X has a continuous distribution, but the proof in the discrete case is virtually identical. Let $d(X) = E(Y|X)$, and let $d^*(X)$ be an arbitrary predictor. We need only prove that $E\{[Y - d(X)]^2\} \le E\{[Y - d^*(X)]^2\}$. It follows from Eq. (4.7.4) that

$$E\{[Y - d(X)]^2\} = E(E\{[Y - d(X)]^2|X\}). \tag{4.7.6}$$

A similar equation holds for d^*. Let $Z = [Y - d(X)]^2$, and let $h(x) = E(Z|x)$. Similarly, let $Z^* = [Y - d^*(X)]^2$ and $h^*(x) = E(Z^*|x)$. The right-hand side of (4.7.6) is $\int h(x) f_1(x)\, dx$, and the corresponding expression using d^* is $\int h^*(x) f_1(x)\, dx$. So, the proof will be complete if we can prove that

$$\int h(x) f_1(x)\, dx \le \int h^*(x) f_1(x)\, dx. \tag{4.7.7}$$

Clearly, Eq. (4.7.7) holds if we can show that $h(x) \le h^*(x)$ for all x. That is, the proof is complete if we can show that $E\{[Y - d(X)]^2|x\} \le E\{[Y - d^*(X)]^2|x\}$. When we condition on $X = x$, we are allowed to treat X as if it were the constant x, so we need to show that $E\{[Y - d(x)]^2|x\} \le E\{[Y - d^*(x)]^2|x\}$. These last expressions are nothing more than the M.S.E.'s for two different predictions $d(x)$ and $d^*(x)$ of Y calculated

using the conditional distribution of Y given $X = x$. As discussed in Sec. 4.5, the M.S.E. of such a prediction is smallest if the prediction is the mean of the distribution of Y. In this case, that mean is the mean of the conditional distribution of Y given $X = x$. Since $d(x)$ is the mean of the conditional distribution of Y given $X = x$, it must have smaller M.S.E. than every other prediction $d^*(x)$. Hence, $h(x) \le h^*(x)$ for all x. ■

If the value $X = x$ is observed and the value $E(Y|x)$ is predicted for Y, then the M.S.E. of this predicted value will be $\mathrm{Var}(Y|x)$, from Definition 4.7.3. It follows from Eq. (4.7.6) that if the prediction is to be made by using the function $d(X) = E(Y|X)$, then the overall M.S.E., averaged over all the possible values of X, will be $E[\mathrm{Var}(Y|X)]$.

If the value of Y must be predicted without any information about the value of X, then, as shown in Sec. 4.5, the best prediction is the mean $E(Y)$ and the M.S.E. is $\mathrm{Var}(Y)$. However, if X can be observed before the prediction is made, the best prediction is $d(X) = E(Y|X)$ and the M.S.E. is $E[\mathrm{Var}(Y|X)]$. Thus, the reduction in the M.S.E. that can be achieved by using the observation X is

$$\mathrm{Var}(Y) - E[\mathrm{Var}(Y|X)]. \tag{4.7.8}$$

This reduction provides a measure of the usefulness of X in predicting Y. It is shown in Exercise 11 at the end of this section that this reduction can also be expressed as $\mathrm{Var}[E(Y|X)]$.

It is important to distinguish carefully between the overall M.S.E., which is $E[\mathrm{Var}(Y|X)]$, and the M.S.E. of the particular prediction to be made when $X = x$, which is $\mathrm{Var}(Y|x)$. *Before* the value of X has been observed, the appropriate value for the M.S.E. of the complete process of observing X and then predicting Y is $E[\mathrm{Var}(Y|X)]$. *After* a particular value x of X has been observed and the prediction $E(Y|x)$ has been made, the appropriate measure of the M.S.E. of this prediction is $\mathrm{Var}(Y|x)$. A useful relationship between these values is given in the following result, whose proof is left to Exercise 11.

Theorem 4.7.4 Law of Total Probability for Variances. If X and Y are arbitrary random variables for which the necessary expectations and variances exist, then $\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}[E(Y|X)]$. ■

Example 4.7.8 A Clinical Trial. In Example 4.7.3, let X be the number of patients out of the first 40 in a clinical trial who have success as their outcome. Let P be the probability that an individual patient is a success. Suppose that P has the uniform distribution on the interval [0, 1] before the trial begins, and suppose that the outcomes of the patients are conditionally independent given $P = p$. As we saw in Example 4.7.3, X has the binomial distribution with parameters 40 and p given $P = p$. If we needed to minimize M.S.E. in predicting P before observing X, we would use the mean of P, namely, 1/2. The M.S.E. would be $\mathrm{Var}(P) = 1/12$. However, we shall soon observe the value of X and then predict P. To do this, we shall need the conditional distribution of P given $X = x$. Bayes' theorem for random variables (3.6.13) tells us that the conditional p.d.f. of P given $X = x$ is

$$g_2(p|x) = \frac{g_1(x|p) f_2(p)}{f_1(x)}, \tag{4.7.9}$$

where $g_1(x|p)$ is the conditional p.f. of X given $P = p$, namely, the binomial p.f. $g_1(x|p) = \binom{40}{x} p^x (1-p)^{40-x}$ for $x = 0, \ldots, 40$, $f_2(p) = 1$ for $0 < p < 1$ is the marginal p.d.f. of P, and $f_1(x)$ is the marginal p.f. of X obtained from the law of total probability

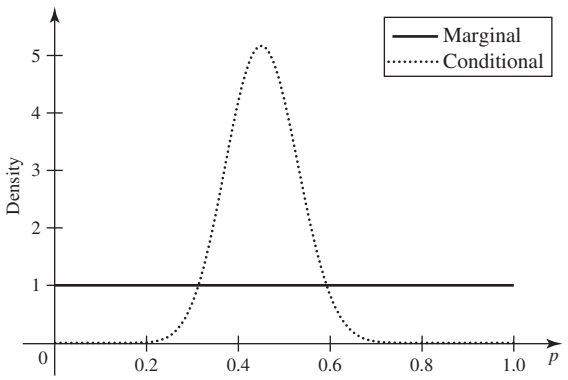

Figure 4.12 The conditional p.d.f. of P given $X = 18$ in Example 4.7.8. The marginal p.d.f. of P (prior to observing X) is also shown.

for random variables (3.6.12):

$$f_1(x) = \int_0^1 \binom{40}{x} p^x (1-p)^{40-x}\, dp. \tag{4.7.10}$$

This last integral looks difficult to compute. However, there is a simple formula for integrals of this form, namely,

$$\int_0^1 p^k (1-p)^\ell\, dp = \frac{k!\ell!}{(k+\ell+1)!}. \tag{4.7.11}$$

A proof of Eq. (4.7.11) is given in Sec. 5.8. Substituting (4.7.11) into (4.7.10) yields

$$f_1(x) = \frac{40!}{x!(40-x)!}\frac{x!(40-x)!}{41!} = \frac{1}{41},$$

for $x = 0, \ldots, 40$. Substituting this into Eq. (4.7.9) yields

$$g_2(p|x) = \frac{41!}{x!(40-x)!} p^x (1-p)^{40-x}, \quad \text{for } 0 < p < 1.$$

For example, with $x = 18$, the observed number of successes in Table 2.1, a graph of $g_2(p|18)$ is shown in Fig. 4.12.

If we want to minimize the M.S.E. when predicting P, we should use $E(P|x)$, the conditional mean. We can compute $E(P|x)$ using the conditional p.d.f. and Eq. (4.7.11):

$$\begin{aligned} E(P|x) &= \int_0^1 p \frac{41!}{x!(40-x)!} p^x (1-p)^{40-x}\, dp \\ &= \frac{41!}{x!(40-x)!}\frac{(x+1)!(40-x)!}{42!} = \frac{x+1}{42}. \end{aligned} \tag{4.7.12}$$

So, after $X = x$ is observed, we will predict P to be $(x+1)/42$, which is very close to the proportion of the first 40 patients who are successes. The M.S.E. after observing $X = x$ is the conditional variance $\mathrm{Var}(P|x)$. We can compute this using (4.7.12) and

$$\begin{aligned} E(P^2|x) &= \int_0^1 p^2 \frac{41!}{x!(40-x)!} p^x (1-p)^{40-x}\, dp \\ &= \frac{41!}{x!(40-x)!}\frac{(x+2)!(40-x)!}{43!} = \frac{(x+1)(x+2)}{42 \times 43}. \end{aligned}$$

Using the fact that $\text{Var}(P|x) = E(P^2|x) - [E(P|x)]^2$, we see that

$$\text{Var}(P|x) = \frac{(x+1)(41-x)}{42^2 \times 43}.$$

The overall M.S.E. of predicting P from X is the mean of the conditional M.S.E.

$$\begin{aligned} E[\text{Var}(P|X)] &= E\left(\frac{(X+1)(41-X)}{42^2 \times 43}\right) \\ &= \frac{1}{75{,}852} E(-X^2 + 40X + 41) \\ &= \frac{1}{75{,}852}\left(-\frac{1}{41}\sum_{x=0}^{40} x^2 + \frac{40}{41}\sum_{x=0}^{40} x + 41\right) \\ &= \frac{1}{75{,}852}\left(-\frac{1}{41}\frac{40 \times 41 \times 81}{6} + \frac{40}{41}\frac{40 \times 41}{2} + 41\right) \\ &= \frac{301}{75{,}852} = 0.003968. \end{aligned}$$

In this calculation, we used two popular formulas,

$$\sum_{k=0}^{n} k = \frac{n(n+1)}{2}, \tag{4.7.13}$$

$$\sum_{k=0}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \tag{4.7.14}$$

The overall M.S.E. is quite a bit smaller than the value $1/12 = 0.08333$, which we would have obtained before observing X. As an illustration, Fig. 4.12 shows how much more spread out the marginal distribution of P is compared to the conditional distribution of P after observing $X = 18$. ◀

It should be emphasized that for the conditions of Example 4.7.8, 0.003968 is the appropriate value of the overall M.S.E. when it is known that the value of X will be available for predicting P but before the explicit value of X has been determined. After the value of $X = x$ has been determined, the appropriate value of the M.S.E. is $\text{Var}(P|x) = \frac{(x+1)(41-x)}{75{,}852}$. Notice that the largest possible value of $\text{Var}(P|x)$ is 0.005814 when $x = 20$ and is still much less than 1/12.

A result similar to Theorem 4.7.3 holds if we are trying to minimize the M.A.E. (mean absolute error) of our prediction rather than the M.S.E. In Exercise 16, you can prove that the predictor that minimizes M.A.E. is $d(X)$ equal to the median of the conditional distribution of Y given X.

Summary

The conditional mean $E(Y|x)$ of Y given $X = x$ is the mean of the conditional distribution of Y given $X = x$. This conditional distribution was defined in Chapter 3. Likewise, the conditional variance $\text{Var}(Y|x)$ of Y given $X = x$ is the variance of the conditional distribution. The law of total probability for expectations says that $E[E(Y|X)] = E(Y)$. If we will observe X and then need to predict Y, the predictor that leads to the smallest M.S.E. is the conditional mean $E(Y|X)$.