Chapter 2

Random Variables and Distributions

CHAPTER OUTLINE

In Chapter 1, we discussed the probability model as the central object of study in the theory of probability. This required defining a probability measure P on a class of subsets of the sample space S. It turns out that there are simpler ways of presenting a particular probability assignment than this — ways that are much more convenient to work with than P. This chapter is concerned with the definitions of random variables, distribution functions, probability functions, density functions, and the development of the concepts necessary for carrying out calculations for a probability model using these entities. This chapter also discusses the concept of the conditional distribution of one random variable, given the values of others. Conditional distributions of random variables provide the framework for discussing what it means to say that variables are related, which is important in many applications of probability and statistics.

2.1 Random Variables

The previous chapter explained how to construct probability models, including a sample space S and a probability measure P. Once we have a probability model, we may define *random variables* for that probability model.

Intuitively, a random variable assigns a numerical value to each possible outcome in the sample space. For example, if the sample space is {rain, snow, clear}, then we might define a random variable X such that $X = 3$ if it rains, $X = 6$ if it snows, and $X = -2.7$ if it is clear.

More formally, we have the following definition.

Definition 2.1.1 A *random variable* is a function from the sample space S to the set R^1 of all real numbers.

Figure 2.1.1 provides a graphical representation of a random variable X taking a response value $s \in S$ into a real number $X(s) \in R^1$.

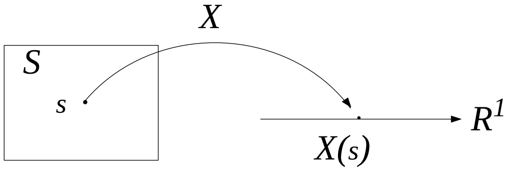

Figure 2.1.1: A random variable X as a function on the sample space S and taking values in R^1.

EXAMPLE 2.1.1 *A Very Simple Random Variable*
The random variable described above could be written formally as X : {rain, snow, clear} $\to R^1$ by $X(\text{rain}) = 3$, $X(\text{snow}) = 6$, and $X(\text{clear}) = -2.7$. We will return to this example below. ∎

We now present several further examples. The point is, we can define random variables any way we like, as long as they are functions from the sample space to R^1.

EXAMPLE 2.1.2
For the case $S =$ {rain, snow, clear}, we might define a second random variable Y by saying that $Y = 0$ if it rains, $Y = -1/2$ if it snows, and $Y = 7/8$ if it is clear. That is $Y(\text{rain}) = 0$, $Y(\text{snow}) = 1/2$, and $Y(\text{clear}) = 7/8$. ∎

EXAMPLE 2.1.3
If the sample space corresponds to flipping three different coins, then we could let X be the total number of heads showing, let Y be the total number of tails showing, let $Z = 0$ if there is exactly one head, and otherwise $Z = 17$, etc. ∎

EXAMPLE 2.1.4
If the sample space corresponds to rolling two fair dice, then we could let X be the square of the number showing on the first die, let Y be the square of the number showing on the second die, let Z be the sum of the two numbers showing, let W be the square of the sum of the two numbers showing, let R be the sum of the squares of the two numbers showing, etc. ∎

EXAMPLE 2.1.5 *Constants as Random Variables*
As a special case, every *constant* value c is also a random variable, by saying that $c(s) = c$ for all $s \in S$. Thus, 5 is a random variable, as is 3 or -21.6. ▮

EXAMPLE 2.1.6 *Indicator Functions*
One special kind of random variable is worth mentioning. If A is any event, then we can define the *indicator function* of A, written I_A, to be the random variable

$$I_A(s) = \begin{cases} 1 & s \in A \\ 0 & s \notin A, \end{cases}$$

which is equal to 1 on A, and is equal to 0 on A^C. ▮

Given random variables X and Y, we can perform the usual arithmetic operations on them. Thus, for example, $Z = X^2$ is another random variable, defined by $Z(s) = X^2(s) = (X(s))^2 = X(s) \times X(s)$. Similarly, if $W = XY^3$, then $W(s) = X(s) \times Y(s) \times Y(s) \times Y(s)$, etc. Also, if $Z = X + Y$, then $Z(s) = X(s) + Y(s)$, etc.

EXAMPLE 2.1.7
Consider rolling a fair six-sided die, so that $S = \{1, 2, 3, 4, 5, 6\}$. Let X be the number showing, so that $X(s) = s$ for $s \in S$. Let Y be three more than the number showing, so that $Y(s) = s + 3$. Let $Z = X^2 + Y$. Then $Z(s) = X(s)^2 + Y(s) = s^2 + s + 3$. So $Z(1) = 5$, $Z(2) = 9$, etc. ▮

We write $X = Y$ to mean that $X(s) = Y(s)$ for all $s \in S$. Similarly, we write $X \le Y$ to mean that $X(s) \le Y(s)$ for all $s \in S$, and $X \ge Y$ to mean that $X(s) \ge Y(s)$ for all $s \in S$. For example, we write $X \le c$ to mean that $X(s) \le c$ for all $s \in S$.

EXAMPLE 2.1.8
Again consider rolling a fair six-sided die, with $S = \{1, 2, 3, 4, 5, 6\}$. For $s \in S$, let $X(s) = s$, and let $Y = X + I_{\{6\}}$. This means that

$$Y(s) = X(s) + I_{\{6\}}(s) = \begin{cases} s & s \le 5 \\ 7 & s = 6. \end{cases}$$

Hence, $Y(s) = X(s)$ for $1 \le s \le 5$. But it is not true that $Y = X$, because $Y(6) \ne X(6)$. On the other hand, it is true that $Y \ge X$. ▮

EXAMPLE 2.1.9
For the random variable of Example 2.1.1 above, it is not true that $X \ge 0$, nor is it true that $X \le 0$. However, it is true that $X \ge -2.7$ and that $X \le 6$. It is also true that $X \ge -10$ and $X \le 100$. ▮

If S is infinite, then a random variable X can take on infinitely many different values.

EXAMPLE 2.1.10
If $S = \{1, 2, 3, \ldots\}$, with $P\{s\} = 2^{-s}$ for all $s \in S$, and if X is defined by $X(s) = s^2$, then we always have $X \ge 1$. But there is no largest value of $X(s)$ because the value $X(s)$ increases without bound as $s \to \infty$. We shall call such a random variable an *unbounded* random variable. ▮

Finally, suppose X is a random variable. We know that different states s occur with different probabilities. It follows that $X(s)$ also takes different values with different probabilities. These probabilities are called the *distribution* of X; we consider them next.

Summary of Section 2.1

- A random variable is a function from the state space to the set of real numbers.
- The function could be constant, or correspond to counting some random quantity that arises, or any other sort of function.

EXERCISES

2.1.1 Let $S = \{1, 2, 3, \ldots\}$, and let $X(s) = s^2$ and $Y(s) = 1/s$ for $s \in S$. For each of the following quantities, determine (with explanation) whether or not it exists. If it does exist, then give its value.
(a) $\min_{s\in S} X(s)$
(b) $\max_{s\in S} X(s)$
(c) $\min_{s\in S} Y(s)$
(d) $\max_{s\in S} Y(s)$

2.1.2 Let $S = \{\text{high, middle, low}\}$. Define random variables X, Y, and Z by $X(\text{high}) = -12$, $X(\text{middle}) = -2$, $X(\text{low}) = 3$, $Y(\text{high}) = 0$, $Y(\text{middle}) = 0$, $Y(\text{low}) = 1$, $Z(\text{high}) = 6$, $Z(\text{middle}) = 0$, $Z(\text{low}) = 4$. Determine whether each of the following relations is true or false.
(a) $X < Y$
(b) $X \le Y$
(c) $Y < Z$
(d) $Y \le Z$
(e) $XY < Z$
(f) $XY \le Z$

2.1.3 Let $S = \{1, 2, 3, 4, 5\}$.
(a) Define two different (i.e., nonequal) nonconstant random variables, X and Y, on S.
(b) For the random variables X and Y that you have chosen, let $Z = X + Y^2$. Compute $Z(s)$ for all $s \in S$.

2.1.4 Consider rolling a fair six-sided die, so that $S = \{1, 2, 3, 4, 5, 6\}$. Let $X(s) = s$, and $Y(s) = s^3 + 2$. Let $Z = XY$. Compute $Z(s)$ for all $s \in S$.

2.1.5 Let A and B be events, and let $X = I_A \cdot I_B$. Is X an indicator function? If yes, then of what event?

2.1.6 Let $S = \{1, 2, 3, 4\}$, $X = I_{\{1,2\}}$, $Y = I_{\{2,3\}}$, and $Z = I_{\{3,4\}}$. Let $W = X + Y + Z$.
(a) Compute $W(1)$.
(b) Compute $W(2)$.
(c) Compute $W(4)$.
(d) Determine whether or not $W \ge Z$.

2.1.7 Let $S = \{1, 2, 3\}$, $X = I_{\{1\}}$, $Y = I_{\{2,3\}}$, and $Z = I_{\{1,2\}}$. Let $W = X - Y + Z$.
(a) Compute $W(1)$.

(b) Compute $W(2)$.
(c) Compute $W(3)$.
(d) Determine whether or not $W \geq Z$.

2.1.8 Let $S = \{1, 2, 3, 4, 5\}$, $X = I_{\{1,2,3\}}$, $Y = I_{\{2,3\}}$, and $Z = I_{\{3,4,5\}}$. Let $W = X - Y + Z$.
(a) Compute $W(1)$.
(b) Compute $W(2)$.
(c) Compute $W(5)$.
(d) Determine whether or not $W \geq Z$.

2.1.9 Let $S = \{1, 2, 3, 4\}$, $X = I_{\{1,2\}}$, and $Y(s) = s^2\, X(s)$.
(a) Compute $Y(1)$.
(b) Compute $Y(2)$.
(c) Compute $Y(4)$.

PROBLEMS

2.1.10 Let X be a random variable.
(a) Is it necessarily true that $X \geq 0$?
(b) Is it necessarily true that there is some real number c such that $X + c \geq 0$?
(c) Suppose the sample space S is finite. Then is it necessarily true that there is some real number c such that $X + c \geq 0$?

2.1.11 Suppose the sample space S is finite. Is it possible to define an unbounded random variable on S? Why or why not?

2.1.12 Suppose X is a random variable that takes only the values 0 or 1. Must X be an indicator function? Explain.

2.1.13 Suppose the sample space S is finite, of size m. How many different indicator functions can be defined on S?

2.1.14 Suppose X is a random variable. Let $Y = \sqrt{X}$. Must Y be a random variable? Explain.

DISCUSSION TOPICS

2.1.15 Mathematical probability theory was introduced to the English-speaking world largely by two American mathematicians, William Feller and Joe Doob, writing in the early 1950s. According to Professor Doob, the two of them had an argument about whether random variables should be called “random variables” or “chance variables.” They decided by flipping a coin — and “random variables” won. (Source: *Statistical Science* **12** (1997), No. 4, page 307.) Which name do *you* think would have been a better choice?

2.2 Distributions of Random Variables

Because random variables are defined to be functions of the outcome s, and because the outcome s is assumed to be random (i.e., to take on different values with different probabilities), it follows that the value of a random variable will itself be random (as the name implies).

Specifically, if X is a random variable, then what is the probability that X will equal some particular value x? Well, $X = x$ precisely when the outcome s is chosen such that $X(s) = x$.

EXAMPLE 2.2.1

Let us again consider the random variable of Example 2.1.1, where $S = \{\text{rain, snow, clear}\}$, and X is defined by $X(\text{rain}) = 3$, $X(\text{snow}) = 6$, and $X(\text{clear}) = -2.7$. Suppose further that the probability measure P is such that $P(\text{rain}) = 0.4$, $P(\text{snow}) = 0.15$, and $P(\text{clear}) = 0.45$. Then clearly, $X = 3$ only when it rains, $X = 6$ only when it snows, and $X = -2.7$ only when it is clear. Thus, $P(X = 3) = P(\text{rain}) = 0.4$, $P(X = 6) = P(\text{snow}) = 0.15$, and $P(X = -2.7) = P(\text{clear}) = 0.45$. Also, $P(X = 17) = 0$, and in fact $P(X = x) = P(\emptyset) = 0$ for all $x \notin \{3, 6, -2.7\}$. We can also compute that

$$P(X \in \{3, 6\}) = P(X = 3) + P(X = 6) = 0.4 + 0.15 = 0.55,$$

while

$$P(X < 5) = P(X = 3) + P(X = -2.7) = 0.4 + 0.45 = 0.85,$$

etc. ▮

We see from this example that, if B is any subset of the real numbers, then $P(X \in B) = P(\{s \in S : X(s) \in B\})$. Furthermore, to understand X well requires knowing the probabilities $P(X \in B)$ for different subsets B. That is the motivation for the following definition.

Definition 2.2.1 If X is a random variable, then the *distribution* of X is the collection of probabilities $P(X \in B)$ for all subsets B of the real numbers.

Strictly speaking, it is required that B be a Borel subset, which is a technical restriction from measure theory that need not concern us here. Any subset that we could ever write down is a Borel subset.

In Figure 2.2.1, we provide a graphical representation of how we compute the distribution of a random variable X. For a set B, we must find the elements in $s \in S$ such that $X(s) \in B$. These elements are given by the set $\{s \in S : X(s) \in B\}$. Then we evaluate the probability $P(\{s \in S : X(s) \in B\})$. We must do this for every subset $B \subset R^1$.

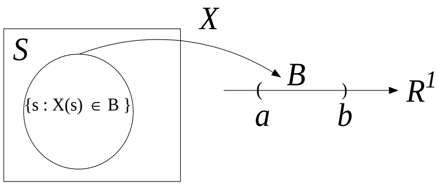

Figure 2.2.1: If $B = (a, b) \subset R^1$, then $\{s \in S : X(s) \in B\}$ is the set of elements such that $a < X(s) < b$.

EXAMPLE 2.2.2 *A Very Simple Distribution*
Consider once again the above random variable, where $S = \{\text{rain, snow, clear}\}$ and where X is defined by $X(\text{rain}) = 3$, $X(\text{snow}) = 6$, and $X(\text{clear}) = -2.7$, and $P(\text{rain}) = 0.4$, $P(\text{snow}) = 0.15$, and $P(\text{clear}) = 0.45$. What is the distribution of X? Well, if B is any subset of the real numbers, then $P(X \in B)$ should count 0.4 if $3 \in B$, plus 0.15 if $6 \in B$, plus 0.45 if $-2.7 \in B$. We can formally write all this information at once by saying that

$$P(X \in B) = 0.4\, I_B(3) + 0.15\, I_B(6) + 0.45\, I_B(-2.7),$$

where again $I_B(x) = 1$ if $x \in B$, and $I_B(x) = 0$ if $x \notin B$. ▮

EXAMPLE 2.2.3 *An Almost-As-Simple Distribution*
Consider once again the above setting, with $S = \{\text{rain, snow, clear}\}$, and $P(\text{rain}) = 0.4$, $P(\text{snow}) = 0.15$, and $P(\text{clear}) = 0.45$. Consider a random variable Y defined by $Y(\text{rain}) = 5$, $Y(\text{snow}) = 7$, and $Y(\text{clear}) = 5$.

What is the distribution of Y? Clearly, $Y = 7$ only when it snows, so that $P(Y = 7) = P(\text{snow}) = 0.15$. However, here $Y = 5$ if it rains *or* if it is clear. Hence, $P(Y = 5) = P(\{\text{rain, clear}\}) = 0.4 + 0.45 = 0.85$. Therefore, if B is any subset of the real numbers, then

$$P(Y \in B) = 0.15\, I_B(7) + 0.85\, I_B(5). \text{ ▮}$$

While the above examples show that it is possible to keep track of $P(X \in B)$ for all subsets B of the real numbers, they also indicate that it is rather cumbersome to do so. Fortunately, there are simpler functions available to help us keep track of probability distributions, including cumulative distribution functions, probability functions, and density functions. We discuss these next.

Summary of Section 2.2

- The distribution of a random variable X is the collection of probabilities $P(X \in B)$ of X belonging to various sets.
- The probability $P(X \in B)$ is determined by calculating the probability of the set of response values s such that $X(s) \in B$, i.e., $P(X \in B) = P(\{s \in S : X(s) \in B\})$.

EXERCISES

2.2.1 Consider flipping two independent fair coins. Let X be the number of heads that appear. Compute $P(X = x)$ for all real numbers x.

2.2.2 Suppose we flip three fair coins, and let X be the number of heads showing.
(a) Compute $P(X = x)$ for every real number x.
(b) Write a formula for $P(X \in B)$, for any subset B of the real numbers.

2.2.3 Suppose we roll two fair six-sided dice, and let Y be the sum of the two numbers showing.
(a) Compute $P(Y = y)$ for every real number y.
(b) Write a formula for $P(Y \in B)$, for any subset B of the real numbers.

2.2.4 Suppose we roll one fair six-sided die, and let Z be the number showing. Let $W = Z^3 + 4$, and let $V = \sqrt{Z}$.
(a) Compute $P(W = w)$ for every real number w.
(b) Compute $P(V = v)$ for every real number v.
(c) Compute $P(ZW = x)$ for every real number x.
(d) Compute $P(VW = y)$ for every real number y.
(e) Compute $P(V + W = r)$ for every real number r.

2.2.5 Suppose that a bowl contains 100 chips: 30 are labelled 1, 20 are labelled 2, and 50 are labelled 3. The chips are thoroughly mixed, a chip is drawn, and the number X on the chip is noted.
(a) Compute $P(X = x)$ for every real number x.
(b) Suppose the first chip is replaced, a second chip is drawn, and the number Y on the chip noted. Compute $P(Y = y)$ for every real number y.
(c) Compute $P(W = w)$ for every real number w when $W = X + Y$.

2.2.6 Suppose a standard deck of 52 playing cards is thoroughly shuffled and a single card is drawn. Suppose an ace has value 1, a jack has value 11, a queen has value 12, and a king has value 13.
(a) Compute $P(X = x)$ for every real number x, when X is the value of the card drawn.
(b) Suppose that $Y = 1, 2, 3,$ or 4 when a diamond, heart, club, or spade is drawn. Compute $P(Y = y)$ for every real number y.
(c) Compute $P(W = w)$ for every real number w when $W = X + Y$.

2.2.7 Suppose a university is composed of 55% female students and 45% male students. A student is selected to complete a questionnaire. There are 25 questions on the questionnaire administered to a male student and 30 questions on the questionnaire administered to a female student. If X denotes the number of questions answered by a randomly selected student, then compute $P(X = x)$ for every real number x.

2.2.8 Suppose that a bowl contains 10 chips, each uniquely numbered 0 through 9. The chips are thoroughly mixed, one is drawn and the number on it, X_1, is noted. This chip is then replaced in the bowl. A second chip is drawn and the number on it, X_2, is noted. Compute $P(W = w)$ for every real number w when $W = X_1 + 10X_2$.

PROBLEMS

2.2.9 Suppose that a bowl contains 10 chips each uniquely numbered 0 through 9. The chips are thoroughly mixed, one is drawn and the number on it, X_1, is noted. This chip is *not* replaced in the bowl. A second chip is drawn and the number on it, X_2, is noted. Compute $P(W = w)$ for every real number w when $W = X_1 + 10X_2$.

CHALLENGES

2.2.10 Suppose Alice flips three fair coins, and let X be the number of heads showing. Suppose Barbara flips five fair coins, and let Y be the number of heads showing. Let $Z = X - Y$. Compute $P(Z = z)$ for every real number z.

2.3 Discrete Distributions

For many random variables X, we have $P(X = x) > 0$ for certain x values. This means there is positive probability that the variable will be equal to certain particular values.

If

$$\sum_{x \in R^1} P(X = x) = 1,$$

then *all* of the probability associated with the random variable X can be found from the probability that X will be equal to certain particular values. This prompts the following definition.

Definition 2.3.1 A random variable X is *discrete* if

$$\sum_{x \in R^1} P(X = x) = 1. \qquad (2.3.1)$$

At first glance one might expect (2.3.1) to be true for *any* random variable. However, (2.3.1) does not hold for the uniform distribution on [0, 1] or for other *continuous* distributions, as we shall see in the next section.

Random variables satisfying (2.3.1) are simple in some sense because we can understand them completely just by understanding their probabilities of being equal to particular values x. Indeed, by simply listing out all the possible values x such that $P(X = x) > 0$, we obtain a second, equivalent definition, as follows.

Definition 2.3.2 A random variable X is *discrete* if there is a finite or countable sequence $x_1, x_2, \ldots$ of distinct real numbers, and a corresponding sequence $p_1, p_2, \ldots$ of nonnegative real numbers, such that $P(X = x_i) = p_i$ for all i, and $\sum_i p_i = 1$.

This second definition also suggests how to keep track of discrete distributions. It prompts the following definition.

Definition 2.3.3 For a discrete random variable X, its *probability function* is the function $p_X : R^1 \to [0, 1]$ defined by

$$p_X(x) = P(X = x).$$

Hence, if $x_1, x_2, \ldots$ are the distinct values such that $P(X = x_i) = p_i$ for all i with $\sum_i p_i = 1$, then

$$p_X(x) = \begin{cases} p_i & x = x_i \text{ for some } i \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, all the information about the distribution of X is contained in its probability function, but only if we *know* that X is a discrete random variable.

Finally, we note that Theorem 1.5.1 immediately implies the following.

Theorem 2.3.1 (*Law of total probability, discrete random variable version*) Let X be a discrete random variable, and let A be some event. Then

$$P(A) = \sum_{x \in R^1} P(X = x)\, P(A \,|\, X = x).$$

2.3.1 Important Discrete Distributions

Certain particular discrete distributions are so important that we list them here.

EXAMPLE 2.3.1 *Degenerate Distributions*
Let c be some fixed real number. Then, as already discussed, c is also a random variable (in fact, c is a *constant random variable*). In this case, clearly c is discrete, with probability function p_c satisfying that $p_c(c) = 1$, and $p_c(x) = 0$ for $x \neq c$. Because c is always equal to a particular value (namely, c) with probability 1, the distribution of c is sometimes called a *point mass* or *point distribution* or *degenerate distribution*. ∎

EXAMPLE 2.3.2 *The Bernoulli Distribution*
Consider flipping a coin that has probability θ of coming up heads and probability $1-\theta$ of coming up tails, where $0 < \theta < 1$. Let $X = 1$ if the coin is heads, while $X = 0$ if the coin is tails. Then $p_X(1) = P(X = 1) = \theta$, while $p_X(0) = P(X = 0) = 1 - \theta$. The random variable X is said to have the Bernoulli(θ) distribution; we write this as $X \sim \text{Bernoulli}(\theta)$.

Bernoulli distributions arise anytime we have a response variable that takes only two possible values, and we label one of these outcomes as 1 and the other as 0. For example, 1 could correspond to success and 0 to failure of some quality test applied to an item produced in a manufacturing process. In this case, θ is the proportion of manufactured items that will pass the test. Alternatively, we could be randomly selecting an individual from a population and recording a 1 when the individual is female and a 0 if the individual is a male. In this case, θ is the proportion of females in the population. ∎

EXAMPLE 2.3.3 *The Binomial Distribution*
Consider flipping n coins, each of which has (independent) probability θ of coming up heads, and probability $1-\theta$ of coming up tails. (Again, $0 < \theta < 1$.) Let X be the total number of heads showing. By (1.4.2), we see that for $x = 0, 1, 2, \ldots, n$,

$$p_X(x) = P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \frac{n!}{x!\,(n-x)!}\,\theta^x (1-\theta)^{n-x}.$$

The random variable X is said to have the Binomial(n, θ) distribution; we write this as $X \sim$ Binomial(n, θ). The Bernoulli(θ) distribution corresponds to the special case of the Binomial(n, θ) distribution when $n = 1$, namely, Bernoulli$(\theta) =$ Binomial$(1, \theta)$. Figure 2.3.1 contains the plots of several Binomial$(20, \theta)$ probability functions.

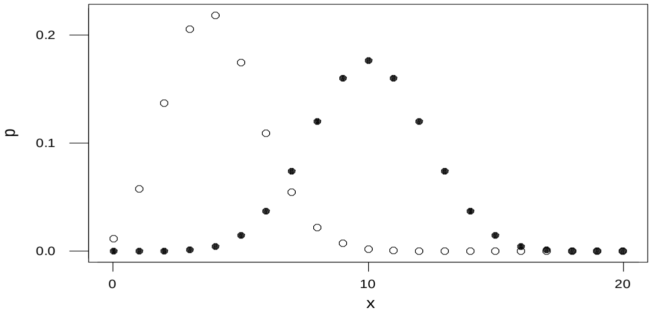

Figure 2.3.1: Plot of the Binomial$(20, 1/2)$ (● ● ●) and the Binomial$(20, 1/5)$ (○ ○ ○) probability functions.

The binomial distribution is applicable to any situation involving n independent performances of a random system; for each performance, we are recording whether a particular event has occurred, called a *success*, or has not occurred, called a *failure*. If we denote the event in question by A and put $\theta = P(A)$, we have that the number of successes in the n performances is distributed Binomial(n, θ). For example, we could be testing light bulbs produced by a manufacturer, and θ is the probability that a bulb works when we test it. Then the number of bulbs that work in a batch of n is distributed Binomial(n, θ). If a baseball player has probability θ of getting a hit when at bat, then the number of hits obtained in n at-bats is distributed Binomial(n, θ).

There is another way of expressing the binomial distribution that is sometimes useful. For example, if $X_1, X_2, \ldots, X_n$ are chosen independently and each has the Bernoulli(θ) distribution, and $Y = X_1 + \cdots + X_n$, then Y will have the Binomial(n, θ) distribution (see Example 3.4.10 for the details). ∎

EXAMPLE 2.3.4 *The Geometric Distribution*
Consider repeatedly flipping a coin that has probability θ of coming up heads and probability $1-\theta$ of coming up tails, where again $0 < \theta < 1$. Let X be the number

of tails that appear before the first head. Then for $k \geq 0$, $X = k$ if and only if the coin shows exactly k tails followed by a head. The probability of this is equal to $(1-\theta)^k\theta$. (In particular, the probability of getting an *infinite* number of tails before the first head is equal to $(1-\theta)^\infty\theta = 0$, so X is never equal to infinity.) Hence, $p_X(k) = (1-\theta)^k\theta$, for $k = 0, 1, 2, 3, \ldots$. The random variable X is said to have the Geometric(θ) distribution; we write this as $X \sim$ Geometric(θ). Figure 2.3.2 contains the plots of several Geometric(θ) probability functions.

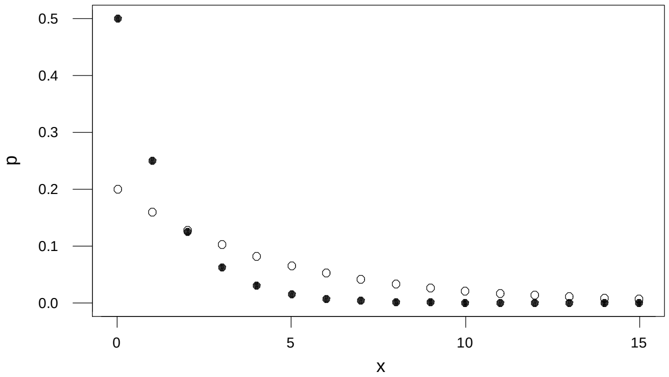

Figure 2.3.2: Plot of the Geometric$(1/2)$ (● ● ●) and the Geometric$(1/5)$ (○ ○ ○) probability functions at the values $0, 1, \ldots, 15$.

The geometric distribution applies whenever we are counting the number of failures until the first success for independent performances of a random system where the occurrence of some event is considered a success. For example, the number of light bulbs tested that work until the first bulb that does not (a working bulb is considered a "failure" for the test) and the number of at-bats without a hit until the first hit for the baseball player both follow the geometric distribution.

We note that some books instead define the geometric distribution to be the number of coin flips up to *and including* the first head, which is simply equal to one plus the random variable defined here. ▮

EXAMPLE 2.3.5 *The Negative-Binomial Distribution*

Generalizing the previous example, consider again repeatedly flipping a coin that has probability θ of coming up heads and probability $1-\theta$ of coming up tails. Let r be a positive integer, and let Y be the number of tails that appear before the rth head. Then for $k \geq 0$, $Y = k$ if and only if the coin shows exactly $r-1$ heads (and k tails) on the first $r-1+k$ flips, and then shows a head on the $(r+k)$-th flip. The probability of this is equal to

$$p_Y(k) = \binom{r-1+k}{r-1}\theta^{r-1}(1-\theta)^k\theta = \binom{r-1+k}{k}\theta^r(1-\theta)^k,$$

for $k = 0, 1, 2, 3, \ldots$.

The random variable Y is said to have the Negative-Binomial(r, θ) distribution; we write this as $Y \sim$ Negative-Binomial(r, θ). Of course, the special case $r = 1$ corresponds to the Geometric(θ) distribution. So in terms of our notation, we have that Negative-Binomial$(1, \theta) =$ Geometric(θ). Figure 2.3.3 contains the plots of several Negative-Binomial(r, θ) probability functions.

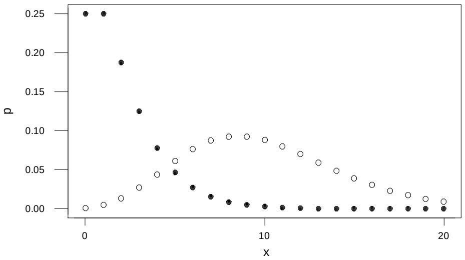

Figure 2.3.3: Plot of the Negative-Binomial$(2, 1/2)$ (● ● ●) probability function and the Negative-Binomial$(10, 1/2)$ (○ ○ ○) probability function at the values $0, 1, \ldots, 20$.

The Negative-Binomial(r, θ) distribution applies whenever we are counting the number of failures until the rth success for independent performances of a random system where the occurrence of some event is considered a success. For example, the number of light bulbs tested that work until the third bulb that does not and the number of at-bats without a hit until the fifth hit for the baseball player both follow the negative-binomial distribution. ▮

EXAMPLE 2.3.6 *The Poisson Distribution*
We say that a random variable Y has the Poisson(λ) distribution, and write $Y \sim$ Poisson(λ), if

$$p_Y(y) = P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

for $y = 0, 1, 2, 3, \ldots$. We note that since (from calculus) $\sum_{y=0}^{\infty} \lambda^y / y! = e^{\lambda}$, it is indeed true (as it must be) that $\sum_{y=0}^{\infty} P(Y = y) = 1$. Figure 2.3.4 contains the plots of several Poisson(λ) probability functions.

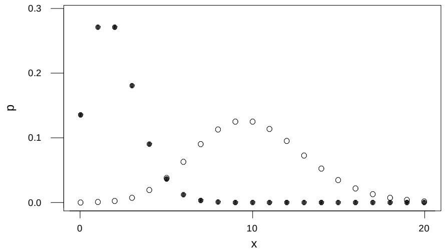

Figure 2.3.4: Plot of the Poisson(2) (● ● ●) and the Poisson(10) (○ ○ ○) probability functions at the values $0, 1, \ldots, 20$.

We motivate the Poisson distribution as follows. Suppose $X \sim \text{Binomial}(n, \theta)$, i.e., X has the Binomial(n, θ) distribution as in Example 2.3.3. Then for $0 \le x \le n$,

$$P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

If we set $\theta = \lambda / n$ for some $\lambda > 0$, then this becomes

$$\begin{aligned} P(X = x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}. \end{aligned} \tag{2.3.2}$$

Let us now consider what happens if we let $n \to \infty$ in (2.3.2), while keeping x fixed at some nonnegative integer. In that case,

$$\frac{n(n-1)\,(n-2)\cdots(n-x+1)}{n^x} = 1\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{x+1}{n}\right)$$

converges to 1 while (since from calculus $(1 + (c/n))^n \to e^c$ for any c)

$$\left(1 - \frac{\lambda}{n}\right)^{n-x} = \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x} \to e^{-\lambda} \cdot 1 = e^{-\lambda}.$$

Substituting these limits into (2.3.2), we see that

$$\lim_{n \to \infty} P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

for $x = 0, 1, 2, 3, \ldots$.

Intuitively, we can phrase this result as follows. If we flip a *very large* number of coins n, and each coin has a *very small* probability $\theta = \lambda/n$ of coming up heads, then the probability that the total number of heads will be x is approximately given by $\lambda^x e^{-\lambda}/x!$. Figure 2.3.5 displays the accuracy of this estimate when we are approximating the Binomial(100, 1/10) distribution by the Poisson(λ) distribution where $\lambda = n\theta = 100(1/10) = 10$.

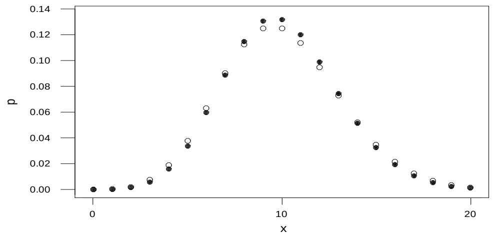

Figure 2.3.5: Plot of the Binomial$(100, 1/10)$ (● ● ●) and the Poisson(10) (○ ○ ○) probability functions at the values $0, 1, \ldots, 20$.

The Poisson distribution is a good model for counting random occurrences of an event when there are many possible occurrences, but each occurrence has very small probability. Examples include the number of house fires in a city on a given day, the number of radioactive events recorded by a Geiger counter, the number of phone calls arriving at a switchboard, the number of hits on a popular World Wide Web page on a given day, etc. ▮

EXAMPLE 2.3.7 *The Hypergeometric Distribution*
Suppose that an urn contains M white balls and $N - M$ black balls. Suppose we draw $n \le N$ balls from the urn in such a fashion that each subset of n balls has the same probability of being drawn. Because there are $\binom{N}{n}$ such subsets, this probability is $1/\binom{N}{n}$.

One way of accomplishing this is to thoroughly mix the balls in the urn and then draw a first ball. Accordingly, each ball has probability $1/N$ of being drawn. Then, without replacing the first ball, we thoroughly mix the balls in the urn and draw a second ball. So each ball in the urn has probability $1/(N-1)$ of being drawn. We then have that any two balls, say the ith and jth balls, have probability

$$\begin{aligned}&P(\text{ball } i \text{ and } j \text{ are drawn})\\&= P(\text{ball } i \text{ is drawn first})P(\text{ball } j \text{ is drawn second} \mid \text{ball } i \text{ is drawn first})\\&\quad + P(\text{ball } j \text{ is drawn first})P(\text{ball } i \text{ is drawn second} \mid \text{ball } j \text{ is drawn first})\\&= \frac{1}{N}\frac{1}{N-1} + \frac{1}{N}\frac{1}{N-1} = 1/\binom{N}{2}\end{aligned}$$

of being drawn in the first two draws. Continuing in this fashion for n draws, we obtain that the probability of any particular set of n balls being drawn is $1/\binom{N}{n}$. This type of sampling is called *sampling without replacement*.

Given that we take a sample of n, let X denote the number of white balls obtained. Note that we must have $X \ge 0$ and $X \ge n - (N - M)$ because at most $N - M$ of the balls could be black. Hence, $X \ge \max(0, n + M - N)$. Furthermore, $X \le n$ and $X \le M$ because there are only M white balls. Hence, $X \le \min(n, M)$.

So suppose $\max(0, n + M - N) \le x \le \min(n, M)$. What is the probability that x white balls are obtained? In other words, what is $P(X = x)$? To evaluate this, we know that we need to count the number of subsets of n balls that contain x white balls. Using the combinatorial principles of Section 1.4.1, we see that this number is given by $\binom{M}{x}\binom{N-M}{n-x}$. Therefore,

$$P(X = x) = \binom{M}{x}\binom{N-M}{n-x} \Big/ \binom{N}{n}$$

for $\max(0, n + M - N) \le x \le \min(n, M)$. The random variable X is said to have the Hypergeometric(N, M, n) distribution. In Figure 2.3.6, we have plotted some hypergeometric probability functions. The Hypergeometric(20, 10, 10) probability function is 0 for $x > 10$, while the Hypergeometric$(20, 10, 5)$ probability function is 0 for $x > 5$.

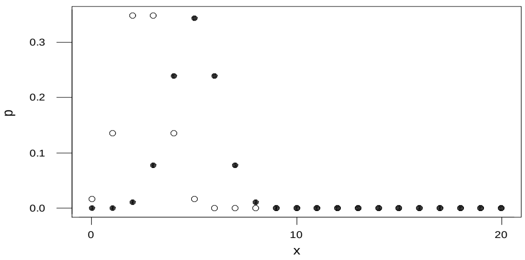

Figure 2.3.6: Plot of Hypergeometric$(20, 10, 10)$ (● ● ●) and Hypergeometric$(20, 10, 5)$ (○ ○ ○) probability functions.

Obviously, the hypergeometric distribution will apply to any context wherein we are *sampling without replacement* from a finite set of N elements and where each element of the set either has a characteristic or does not. For example, if we randomly select people to participate in an opinion poll so that each set of n individuals in a population of N has the same probability of being selected, then the number of people who respond yes to a particular question is distributed Hypergeometric(N, M, n), where M is the number of people in the entire population who would respond yes. We will see the relevance of this to statistics in Section 5.4.2. ▮

Suppose in Example 2.3.7 we had instead *replaced* the drawn ball before drawing the next ball. This is called *sampling with replacement*. It is then clear, from Example 2.3.3, that the number of white balls observed in n draws is distributed Binomial$(n, M/N)$.

Summary of Section 2.3

- A random variable X is discrete if $\sum_x P(X = x) = 1$, i.e., if all its probability comes from being equal to particular values.
- A discrete random variable X takes on only a finite, or countable, number of distinct values.
- Important discrete distributions include the degenerate, Bernoulli, binomial, geometric, negative-binomial, Poisson, and hypergeometric distributions.

EXERCISES

2.3.1 Consider rolling two fair six-sided dice. Let Y be the sum of the numbers showing. What is the probability function of Y?

2.3.2 Consider flipping a fair coin. Let $Z = 1$ if the coin is heads, and $Z = 3$ if the coin is tails. Let $W = Z^2 + Z$.
(a) What is the probability function of Z?
(b) What is the probability function of W?

2.3.3 Consider flipping two fair coins. Let $X = 1$ if the first coin is heads, and $X = 0$ if the first coin is tails. Let $Y = 1$ if the second coin is heads, and $Y = 5$ if the second coin is tails. Let $Z = XY$. What is the probability function of Z?

2.3.4 Consider flipping two fair coins. Let $X = 1$ if the first coin is heads, and $X = 0$ if the first coin is tails. Let $Y = 1$ if the two coins show the *same* thing (i.e., both heads or both tails), with $Y = 0$ otherwise. Let $Z = X + Y$, and $W = XY$.
(a) What is the probability function of Z?
(b) What is the probability function of W?

2.3.5 Consider rolling two fair six-sided dice. Let W be the product of the numbers showing. What is the probability function of W?

2.3.6 Let $Z \sim \text{Geometric}(\theta)$. Compute $P(5 \le Z \le 9)$.

2.3.7 Let $X \sim \text{Binomial}(12, \theta)$. For what value of θ is $P(X = 11)$ maximized?

2.3.8 Let $W \sim \text{Poisson}(\lambda)$. For what value of λ is $P(W = 11)$ maximized?

2.3.9 Let $Z \sim \text{Negative-Binomial}(3, 1/4)$. Compute $P(Z \le 2)$.

2.3.10 Let $X \sim \text{Geometric}(1/5)$. Compute $P(X^2 \le 15)$.

2.3.11 Let $Y \sim \text{Binomial}(10, \theta)$. Compute $P(Y = 10)$.

2.3.12 Let $X \sim \text{Poisson}(\lambda)$. Let $Y = X - 7$. What is the probability function of Y?

2.3.13 Let $X \sim \text{Hypergeometric}(20, 7, 8)$. What is the probability that $X = 3$? What is the probability that $X = 8$?

2.3.14 Suppose that a symmetrical die is rolled 20 independent times, and each time we record whether or not the event $\{2, 3, 5, 6\}$ has occurred.

(a) What is the distribution of the number of times this event occurs in 20 rolls?

(b) Calculate the probability that the event occurs five times.

2.3.15 Suppose that a basketball player sinks a basket from a certain position on the court with probability 0.35.

(a) What is the probability that the player sinks three baskets in 10 independent throws?

(b) What is the probability that the player throws 10 times before obtaining the first basket?

(c) What is the probability that the player throws 10 times before obtaining two baskets?

2.3.16 An urn contains 4 black balls and 5 white balls. After a thorough mixing, a ball is drawn from the urn, its color is noted, and the ball is returned to the urn.

(a) What is the probability that 5 black balls are observed in 15 such draws?

(b) What is the probability that 15 draws are required until the first black ball is observed?

(c) What is the probability that 15 draws are required until the fifth black ball is observed?

2.3.17 An urn contains 4 black balls and 5 white balls. After a thorough mixing, a ball is drawn from the urn, its color is noted, and the ball is set aside. The remaining balls are then mixed and a second ball is drawn.

(a) What is the probability distribution of the number of black balls observed?

(b) What is the probability distribution of the number of white balls observed?

2.3.18 (*Poisson processes and queues*) Consider a situation involving a server, e.g., a cashier at a fast-food restaurant, an automatic bank teller machine, a telephone exchange, etc. Units typically arrive for service in a random fashion and form a queue when the server is busy. It is often the case that the number of arrivals at the server, for some specific unit of time t, can be modeled by a Poisson(λt) distribution and is such that the number of arrivals in nonoverlapping periods are independent. In Chapter 3, we will show that λt is the average number of arrivals during a time period of length t, and so λ is the rate of arrivals per unit of time.

Suppose telephone calls arrive at a help line at the rate of two per minute. A Poisson process provides a good model.

(a) What is the probability that five calls arrive in the next 2 minutes?

(b) What is the probability that five calls arrive in the next 2 minutes and then five more calls arrive in the following 2 minutes?

(c) What is the probability that no calls will arrive during a 10-minute period?

2.3.19 Suppose an urn contains 1000 balls — one of these is black, and the other 999 are white. Suppose that 100 balls are randomly drawn from the urn with replacement. Use the appropriate Poisson distribution to approximate the probability that five black balls are observed.

2.3.20 Suppose that there is a loop in a computer program and that the test to exit the loop depends on the value of a random variable X. The program exits the loop whenever $X \in A$, and this occurs with probability 1/3. If the loop is executed at least once, what is the probability that the loop is executed five times before exiting?

COMPUTER EXERCISES

2.3.21 Tabulate and plot the Hypergeometric(20, 8, 10) probability function.

2.3.22 Tabulate and plot the Binomial(30, 0.3) probability function. Tabulate and plot the Binomial(30, 0.7) probability function. Explain why the Binomial(30, 0.3) probability function at x agrees with the Binomial(30, 0.7) probability function at $n - x$.

PROBLEMS

2.3.23 Let X be a discrete random variable with probability function $p_X(x) = 2^{-x}$ for $x = 1, 2, 3, \ldots$, with $p_X(x) = 0$ otherwise.
(a) Let $Y = X^2$. What is the probability function p_Y of Y?
(b) Let $Z = X - 1$. What is the distribution of Z? (Identify the distribution by name and specify all parameter values.)

2.3.24 Let $X \sim$ Binomial(n_1, θ) and $Y \sim$ Binomial(n_2, θ), with X and Y chosen independently. Let $Z = X + Y$. What will be the distribution of Z? (Explain your reasoning.) (Hint: See the end of Example 2.3.3.)

2.3.25 Let $X \sim$ Geometric(θ) and $Y \sim$ Geometric(θ), with X and Y chosen independently. Let $Z = X + Y$. What will be the distribution of Z? Generalize this to r coins. (Explain your reasoning.)

2.3.26 Let $X \sim \text{Geometric}(\theta_1)$ and $Y \sim \text{Geometric}(\theta_2)$, with X and Y chosen independently. Compute $P(X \le Y)$. Explain what this probability is in terms of coin tossing.

2.3.27 Suppose that $X \sim \text{Geometric}(\lambda/n)$. Compute $\lim_{n\to\infty} P(X \le n)$.

2.3.28 Let $X \sim \text{Negative-Binomial}(r, \theta)$ and $Y \sim \text{Negative-Binomial}(s, \theta)$, with X and Y chosen independently. Let $Z = X + Y$. What will be the distribution of Z? (Explain your reasoning.)

2.3.29 (*Generalized hypergeometric distribution*) Suppose that a set contains N objects, M_1 of which are labelled 1, M_2 of which are labelled 2, and the remainder of which are labelled 3. Suppose we select a sample of $n \le N$ objects from the set using sampling without replacement, as described in Example 2.3.7. Determine the probability that we obtain the counts (f_1, f_2, f_3) where f_i is the number of objects labelled i in the sample.

2.3.30 Suppose that units arrive at a server according to a Poisson process at rate λ (see Exercise 2.3.18). Let T be the amount of time until the first call. Calculate $P(T > t)$.

2.4 Continuous Distributions

In the previous section, we considered discrete random variables X for which $P(X = x) > 0$ for certain values of x. However, for some random variables X, such as one having the uniform distribution, we have $P(X = x) = 0$ for all x. This prompts the following definition.

Definition 2.4.1 A random variable X is *continuous* if

$$P(X = x) = 0\,, \tag{2.4.1}$$

for all $x \in R^1$.

EXAMPLE 2.4.1 *The Uniform*[0, 1] *Distribution*

Consider a random variable whose distribution is the uniform distribution on [0, 1], as presented in (1.2.2). That is,

$$P(a \le X \le b) = b - a, \tag{2.4.2}$$

whenever $0 \le a \le b \le 1$, with $P(X < 0) = P(X > 1) = 0$. The random variable X is said to have the Uniform[0, 1] distribution; we write this as $X \sim \text{Uniform}[0, 1]$. For example,

$$P\left(\frac{1}{2} \le X \le \frac{3}{4}\right) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}.$$

Also,

$$P\left(X \ge \frac{2}{3}\right) = P\left(\frac{2}{3} \le X \le 1\right) + P\,(X > 1)) = \left(1 - \frac{2}{3}\right) + 0 = \frac{1}{3}.$$

In fact, for any $x \in [0, 1]$,

$$P(X \le x) = P(X < 0) + P(0 \le X \le x) = 0 + (x - 0) = x.$$

Note that setting $a = b = x$ in (2.4.2), we see in particular that $P(X = x) = x - x = 0$ for every $x \in R^1$. Thus, the uniform distribution is an example of a continuous distribution. In fact, it is one of the most important examples! ∎

The Uniform[0, 1] distribution is fairly easy to work with. However, in general, continuous distributions are very difficult to work with. Because $P(X = x) = 0$ for all x, we cannot simply add up probabilities as we can for discrete random variables. Thus, how can we keep track of all the probabilities?

A possible solution is suggested by rewriting (2.4.2), as follows. For $x \in R^1$, let

$$f(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.4.3}$$

Then (2.4.2) can be rewritten as

$$P(a \le X \le b) = \int_a^b f(x)\, dx, \tag{2.4.4}$$

whenever $a \le b$.

One might wonder about the wisdom of converting the simple equation (2.4.2) into the complicated integral equation (2.4.4). However, the advantage of (2.4.4) is that, by modifying the function f, we can obtain many other continuous distributions besides the uniform distribution. To explore this, we make the following definitions.

Definition 2.4.2 Let $f : R^1 \to R^1$ be a function. Then f is a *density function* if $f(x) \ge 0$ for all $x \in R^1$, and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

Definition 2.4.3 A random variable X is *absolutely continuous* if there is a density function f, such that

$$P(a \le X \le b) = \int_a^b f(x)\, dx, \tag{2.4.5}$$

whenever $a \le b$, as in (2.4.4).

In particular, if $b = a + \delta$, with δ a small positive number, and if f is continuous at a, then we see that

$$P(a \le X \le a + \delta) = \int_a^{a+\delta} f(x)\, dx \approx \delta\, f(a).$$

Thus, a density function evaluated at a may be thought of as measuring the probability of a random variable being in a small interval about a.

To better understand absolutely continuous random variables, we note the following theorem.

Theorem 2.4.1 Let X be an absolutely continuous random variable. Then X is a continuous random variable, i.e., $P(X = a) = 0$ for all $a \in R^1$.

PROOF Let a be any real number. Then $P(X = a) = P(a \le X \le a)$. On the other hand, setting $a = b$ in (2.4.5), we see that $P(a \le X \le a) = \int_a^a f(x)\,dx = 0$. Hence, $P(X = a) = 0$ for all a, as required. ∎

It turns out that the converse to Theorem 2.4.1 is false. That is, not all continuous distributions are absolutely continuous.[1] However, most of the continuous distributions that arise in statistics are absolutely continuous. Furthermore, absolutely continuous distributions are much easier to work with than are other kinds of continuous distributions. Hence, we restrict our discussion to absolutely continuous distributions here. In fact, statisticians sometimes say that X is *continuous* as shorthand for saying that X is absolutely continuous.

2.4.1 Important Absolutely Continuous Distributions

Certain absolutely continuous distributions are so important that we list them here.

EXAMPLE 2.4.2 *The Uniform*[0, 1] *Distribution*
Clearly, the uniform distribution is absolutely continuous, with the density function given by (2.4.3). We will see, in Section 2.10, that the Uniform[0, 1] distribution has an important relationship with every absolutely continuous distribution. ∎

EXAMPLE 2.4.3 *The Uniform*$[L, R]$ *Distribution*
Let L and R be any two real numbers with $L < R$. Consider a random variable X such that

$$P(a \le X \le b) = \frac{b - a}{R - L} \tag{2.4.6}$$

whenever $L \le a \le b \le R$, with $P(X < L) = P(X > R) = 0$. The random variable X is said to have the Uniform$[L, R]$ distribution; we write this as $X \sim$ Uniform$[L, R]$. (If $L = 0$ and $R = 1$, then this definition coincides with the previous definition of the Uniform[0, 1] distribution.) Note that $X \sim$ Uniform$[L, R]$ has the same probability of being in any two subintervals of $[L, R]$ that have the same length.

Note that the Uniform$[L, R]$ distribution is also absolutely continuous, with density given by

$$f(x) = \begin{cases} \frac{1}{R-L} & L \le x \le R \\ 0 & \text{otherwise.} \end{cases}$$

In Figure 2.4.1 we have plotted a Uniform[2, 4] density. ∎

[1] For examples of this, see more advanced probability books, e.g., page 143 of *A First Look at Rigorous Probability Theory,* Second Edition, by J. S. Rosenthal (World Scientific Publishing, Singapore, 2006).

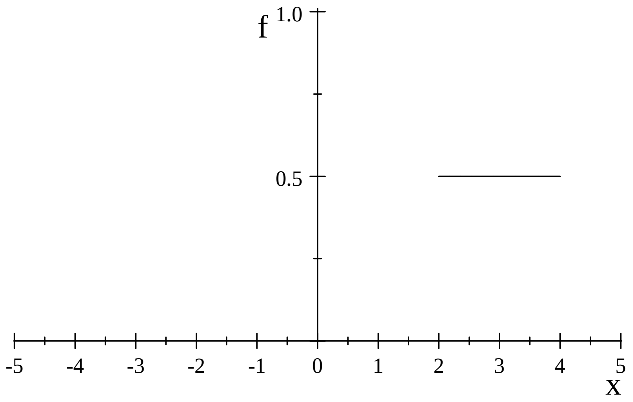

Figure 2.4.1: A Uniform$[2, 4]$ density function.

EXAMPLE 2.4.4 *The Exponential*(1) *Distribution*
Define a function $f : R^1 \to R^1$ by

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

Then clearly, $f(x) \geq 0$ for all x. Also,

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} e^{-x}\,dx = -e^{-x}\ \Big|_0^{\infty} = (-0) - (-1) = 1\,.$$

Hence, f is a density function. See Figure 2.4.2 for a plot of this density.

Consider now a random variable X having this density function f. If $0 \leq a \leq b < \infty$, then

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

The random variable X is said to have the Exponential(1) distribution, which we write as $X \sim$ Exponential(1). The exponential distribution has many important properties, which we will explore in the coming sections. ∎

EXAMPLE 2.4.5 *The Exponential*(λ) *Distribution*
Let $\lambda > 0$ be a fixed constant. Define a function $f : R^1 \to R^1$ by

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

Then clearly, $f(x) \geq 0$ for all x. Also,

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\ \Big|_0^{\infty} = (-0) - (-1) = 1.$$

Hence, f is again a density function. (If $\lambda = 1$, then this corresponds to the Exponential(1) density.)

If X is a random variable having this density function f, then

$$P(a \le X \le b) = \int_a^b \lambda\, e^{-\lambda x}\, dx = (-e^{-\lambda b}) - (-e^{-\lambda a}) = e^{-\lambda a} - e^{-\lambda b}$$

for $0 \le a \le b < \infty$. The random variable X is said to have the Exponential(λ) distribution; we write this as $X \sim$ Exponential(λ). Note that some books and software packages instead replace λ by $1/\lambda$ in the definition of the Exponential(λ) distribution — always check this when using another book or when using software.

An exponential distribution can often be used to model lifelengths. For example, a certain type of light bulb produced by a manufacturer might follow an Exponential(λ) distribution for an appropriate choice of λ. By this we mean that the lifelength X of a randomly selected light bulb from those produced by this manufacturer has probability

$$P(X \ge x) = \int_x^\infty \lambda e^{-\lambda z}\, dz = e^{-\lambda x}$$

of lasting longer than x, in whatever units of time are being used. We will see in Chapter 3 that, in a specific application, the value $1/\lambda$ will correspond to the average lifelength of the light bulbs.

As another application of this distribution, consider a situation involving a server, e.g., a cashier at a fast-food restaurant, an automatic bank teller machine, a telephone exchange, etc. Units arrive for service in a random fashion and form a queue when the server is busy. It is often the case that the number of arrivals at the server, for some specific unit of time t, can be modeled by a Poisson(λt) distribution. Now let T_1 be the time until the first arrival. Then we have

$$P\,(T_1 > t) = P\,(\text{no arrivals in } (0,t]) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

and T_1 has density given by

$$f\,(t) = -\frac{d}{dt}\int_t^\infty f\,(z)\; dz = -\frac{d}{dt} P\,(T_1 > t) = \lambda e^{-\lambda t}.$$

So $T_1 \sim$ Exponential(λ). ▮

EXAMPLE 2.4.6 *The Gamma(α, λ) Distribution*
The *gamma function* is defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt\,, \qquad \alpha > 0.$$

It turns out (see Problem 2.4.15) that

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \tag{2.4.7}$$

and that if n is a positive integer, then $\Gamma(n) = (n-1)!$, while $\Gamma(1/2) = \sqrt{\pi}$.

We can use the gamma function to define the density of the Gamma(α, λ) distribution, as follows. Let $\alpha > 0$ and $\lambda > 0$, and define a function f by

$$f(x) = \frac{\lambda^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} \tag{2.4.8}$$

when $x > 0$, with $f(x) = 0$ for $x \leq 0$. Then clearly $f \geq 0$. Furthermore, it is not hard to verify (see Problem 2.4.17) that $\int_0^\infty f(x)\, dx = 1$. Hence, f is a density function.

A random variable X having density function f given by (2.4.8) is said to have the Gamma(α, λ) distribution; we write this as $X \sim$ Gamma(α, λ). Note that some books and software packages instead replace λ by $1/\lambda$ in the definition of the Gamma(α, λ) distribution — always check this when using another book or when using software.

The case $\alpha = 1$ corresponds (because $\Gamma(1) = 0! = 1$) to the Exponential(λ) distribution: Gamma$(1, \lambda) =$ Exponential(λ). In Figure 2.4.2, we have plotted several Gamma(α, λ) density functions.

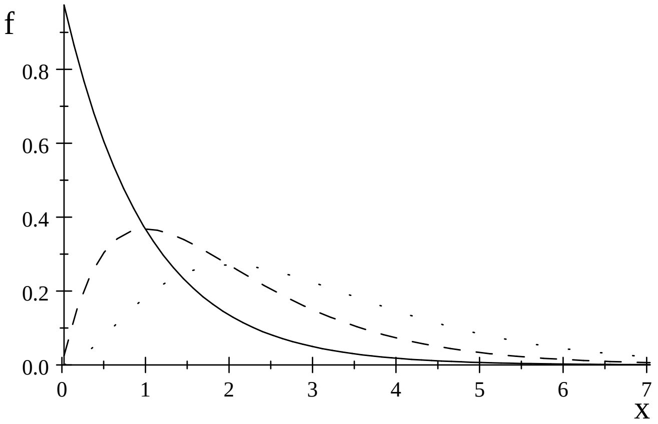

Figure 2.4.2: Graph of an Exponential(1) (solid line), a Gamma$(2, 1)$ (dashed line), and a Gamma$(3, 1)$ (dotted line) density.

A gamma distribution can also be used to model lifelengths. As Figure 2.4.2 shows, the gamma family gives a much greater variety of shapes to choose from than from the exponential family. ▮

We now define a function $\phi : R^1 \to R^1$ by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{2.4.9}$$

This function ϕ is the famous "bell-shaped curve" because its graph is in the shape of a bell, as shown in Figure 2.4.3.

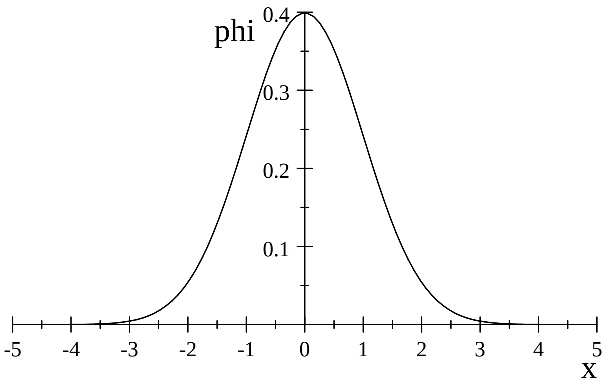

Figure 2.4.3: Plot of the function ϕ in (2.4.9).

We have the following result for ϕ.

Theorem 2.4.2 The function ϕ given by (2.4.9) is a density function.

PROOF See Section 2.11 for the proof of this result. ▮

This leads to the following important distributions.

EXAMPLE 2.4.7 *The $N(0, 1)$ Distribution*
Let X be a random variable having the density function ϕ given by (2.4.9). This means that for $-\infty < a \le b < \infty$,

$$P(a \le X \le b) = \int_a^b \phi(x)\,dx = \int_a^b \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}\,dx.$$

The random variable X is said to have the $N(0, 1)$ distribution (or the *standard normal distribution*); we write this as $X \sim N(0, 1)$. ▮

EXAMPLE 2.4.8 *The $N(\mu, \sigma^2)$ Distribution*
Let $\mu \in R^1$, and let $\sigma > 0$. Let f be the function defined by

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

(If $\mu = 0$ and $\sigma = 1$, then this corresponds with the previous example.) Clearly, $f \ge 0$. Also, letting $y = (x - \mu)/\sigma$, we have

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

Hence, f is a density function.

Let X be a random variable having this density function f. The random variable X is said to have the $N(\mu, \sigma^2)$ distribution; we write this as $X \sim N(\mu, \sigma^2)$. In Figure 2.4.4, we have plotted the $N(0, 1)$ and the $N(1, 1)$ densities. Note that changes in μ

simply shift the density without changing its shape. In Figure 2.4.5, we have plotted the $N(0, 1)$ and the $N(0, 4)$ densities. Note that both densities are centered on 0, but the $N(0, 4)$ density is much more spread out. The value of σ^2 controls the amount of spread.

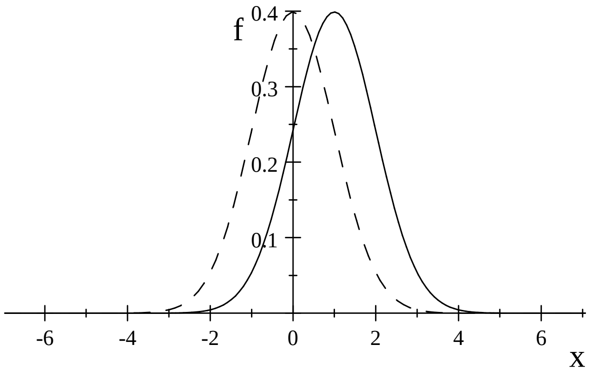

Figure 2.4.4: Graph of the $N(1, 1)$ density (solid line) and the $N(0, 1)$ density (dashed line).

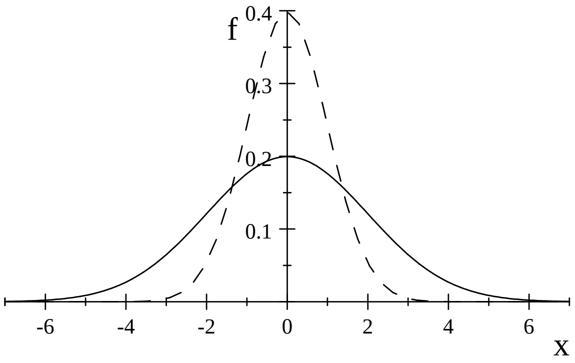

Figure 2.4.5: Graph of an $N(0, 4)$ density (solid line) and an $N(0, 1)$ density (dashed line).

The $N(\mu, \sigma^2)$ distribution, for some choice of μ and σ^2, arises quite often in applications. Part of the reason for this is an important result known as the central limit theorem. which we will discuss in Section 4.4. In particular, this result leads to using a normal distribution to approximate other distributions, just as we used the Poisson distribution to approximate the binomial distribution in Example 2.3.6.

In a large human population, it is not uncommon for various body measurements to be normally distributed (at least to a reasonable degree of approximation). For example, let us suppose that heights (measured in feet) of students at a particular university are distributed $N(\mu, \sigma^2)$ for some choice of μ and σ^2. Then the probability that a randomly selected student has height between a and b feet, with $a < b$, is given by

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx.$$

In Section 2.5, we will discuss how to evaluate such an integral. Later in this text, we will discuss how to select an appropriate value for μ and σ^2 and to assess whether or not any normal distribution is appropriate to model the distribution of a variable defined on a particular population. ▮

Given an absolutely continuous random variable X, we will write its density as f_X, or as f if no confusion arises. Absolutely continuous random variables will be used extensively in later chapters of this book.

Remark 2.4.1 Finally, we note that density functions are not unique. Indeed, if f is a density function and we change its value at a finite number of points, then the value of $\int_a^b f(x)\,dx$ will remain unchanged. Hence, the changed function will also qualify as a density corresponding to the same distribution. On the other hand, often a particular "best" choice of density function is clear. For example, if the density function can be chosen to be continuous, or even piecewise continuous, then this is preferred over some other version of the density function.

To take a specific example, for the Uniform[0, 1] distribution, we could replace the density f of (2.4.3) by

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

or even by

$$h(x) = \begin{cases} 1 & 0 < x < 3/4 \\ 17 & x = 3/4 \\ 1 & 3/4 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Either of these new densities would again define the Uniform[0, 1] distribution, because we would have $\int_a^b f(x)\,dx = \int_a^b g(x)\,dx = \int_a^b h(x)\,dx$ for any $a < b$.

On the other hand, the densities f and g are both piecewise continuous and are therefore natural choices for the density function, whereas h is an unnecessarily complicated choice. Hence, when dealing with density functions, we shall always assume that they are as continuous as possible, such as f and g, rather than having removable discontinuities such as h. This will be particularly important when discussing likelihood methods in Chapter 6. ▮

Summary of Section 2.4

- A random variable X is continuous if $P(X = x) = 0$ for all x, i.e., if none of its probability comes from being equal to particular values.
- X is absolutely continuous if there exists a density function f_X with $P(a \le X \le b) = \int_a^b f_X(x)\,dx$ for all $a < b$.
- Important absolutely continuous distributions include the uniform, exponential, gamma, and normal.

EXERCISES

2.4.1 Let $U \sim \text{Uniform}[0, 1]$. Compute each of the following.
(a) $P(U \le 0)$
(b) $P(U = 1/2)$
(c) $P(U < -1/3)$
(d) $P(U \le 2/3)$
(e) $P(U < 2/3)$
(f) $P(U < 1)$
(g) $P(U \le 17)$

2.4.2 Let $W \sim \text{Uniform}[1, 4]$. Compute each of the following.
(a) $P(W \ge 5)$
(b) $P(W \ge 2)$
(c) $P(W^2 \le 9)$ (Hint: If $W^2 \le 9$, what must W be?)
(d) $P(W^2 \le 2)$

2.4.3 Let $Z \sim \text{Exponential}(4)$. Compute each of the following.
(a) $P(Z \ge 5)$
(b) $P(Z \ge -5)$
(c) $P(Z^2 \ge 9)$
(d) $P(Z^4 - 17 \ge 9)$

2.4.4 Establish for which constants c the following functions are densities.

(a) $f(x) = cx$ on $(0, 1)$ and 0 otherwise.

(b) $f(x) = cx^n$ on $(0, 1)$ and 0 otherwise, for n a nonnegative integer.

(c) $f(x) = cx^{1/2}$ on $(0, 2)$ and 0 otherwise.

(d) $f(x) = c \sin x$ on $(0, \pi/2)$ and 0 otherwise.

2.4.5 Is the function defined by $f(x) = x/3$ for $-1 < x < 2$ and 0 otherwise, a density? Why or why not?

2.4.6 Let $X \sim \text{Exponential}(3)$. Compute each of the following.
(a) $P(0 < X < 1)$
(b) $P(0 < X < 3)$
(c) $P(0 < X < 5)$
(d) $P(2 < X < 5)$
(e) $P(2 < X < 10)$
(f) $P(X > 2)$

2.4.7 Let $M > 0$, and suppose $f(x) = cx^2$ for $0 < x < M$, otherwise $f(x) = 0$. For what value of c (depending on M) is f a density?

2.4.8 Suppose X has density f and that $f(x) \ge 2$ for $0.3 < x < 0.4$. Prove that $P(0.3 < X < 0.4) \ge 0.2$.

2.4.9 Suppose X has density f and Y has density g. Suppose $f(x) > g(x)$ for $1 < x < 2$. Prove that $P(1 < X < 2) > P(1 < Y < 2)$.

2.4.10 Suppose X has density f and Y has density g. Is it possible that $f(x) > g(x)$ for all x? Explain.

2.4.11 Suppose X has density f and $f(x) > f(y)$ whenever $0 < x < 1 < y < 2$. Does it follow that $P(0 < X < 1) > P(1 < X < 2)$? Explain.

2.4.12 Suppose X has density f and $f(x) > f(y)$ whenever $0 < x < 1 < y < 3$. Does it follow that $P(0 < X < 1) > P(1 < X < 3)$? Explain.

2.4.13 Suppose $X \sim N(0, 1)$ and $Y \sim N(1, 1)$. Prove that $P(X < 3) > P(Y < 3)$.

PROBLEMS

2.4.14 Let $Y \sim \text{Exponential}(\lambda)$ for some $\lambda > 0$. Let $y, h \geq 0$. Prove that $P(Y - h \geq y \,|\, Y \geq h) = P(Y \geq y)$. That is, conditional on knowing that $Y \geq h$, the random variable $Y - h$ has the same distribution as Y did originally. This is called the *memoryless* property of the exponential distributions; it says that they immediately "forget" their past behavior.

2.4.15 Consider the gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$, for $\alpha > 0$.
(a) Prove that $\Gamma(\alpha + 1) = \alpha\,\Gamma(\alpha)$. (Hint: Use integration by parts.)
(b) Prove that $\Gamma(1) = 1$.
(c) Use parts (a) and (b) to show that $\Gamma(n) = (n - 1)!$ if n is a positive integer.

2.4.16 Use the fact that $\Gamma(1/2) = \sqrt{\pi}$ to give an alternate proof that $\int_{-\infty}^{\infty} \phi(x)\, dx$ $= 1$ (as in Theorem 2.4.2). (Hint: Make the substitution $t = x^2/2$.)

2.4.17 Let f be the density of the Gamma(α, λ) distribution, as in (2.4.8). Prove that $\int_0^\infty f(x)\, dx = 1$. (Hint: Let $t = \lambda x$.)

2.4.18 (*Logistic distribution*) Consider the function given by $f(x) =$ $e^{-x}\left(1 + e^{-x}\right)^{-2}$ for $-\infty < x < \infty$. Prove that f is a density function.

2.4.19 (*Weibull*(α) *distribution*) Consider, for $\alpha > 0$ fixed, the function given by $f(x) = \alpha x^{\alpha-1}e^{-x^\alpha}$ for $0 < x < \infty$ and 0 otherwise. Prove that f is a density function.

2.4.20 (*Pareto*(α) *distribution*) Consider, for $\alpha > 0$ fixed, the function given by $f(x) = \alpha\,(1 + x)^{-\alpha-1}$ for $0 < x < \infty$ and 0 otherwise. Prove that f is a density function.

2.4.21 (*Cauchy distribution*) Consider the function given by

$$f(x) = \frac{1}{\pi}\frac{1}{1 + x^2}$$

for $-\infty < x < \infty$. Prove that f is a density function. (Hint: Recall the derivative of $\arctan(x)$.)

2.4.22 (*Laplace distribution*) Consider the function given by $f(x) = e^{-|x|}/2$ for $-\infty < x < \infty$ and 0 otherwise. Prove that f is a density function.

2.4.23 (*Extreme value distribution*) Consider the function given by $f(x) = e^{-x} \exp\left\{-e^{-x}\right\}$ for $-\infty < x < \infty$ and 0 otherwise. Prove that f is a density function.

2.4.24 (*Beta*(a, b) *distribution*) The *beta function* is the function $B : (0, \infty)^2 \to R^1$ given by

$$B\,(a, b) = \int_0^1 x^{a-1}\,(1 - x)^{b-1}\; dx.$$

It can be proved (see Challenge 2.4.25) that

$$B\,(a, b) = \frac{\Gamma\,(a)\,\Gamma\,(b)}{\Gamma\,(a + b)} \tag{2.4.10}$$

(a) Prove that the function f given by $f(x) = B^{-1}(a, b)\, x^{a-1}(1-x)^{b-1}$, for $0 < x < 1$ and 0 otherwise, is a density function.

(b) Determine and plot the density when $a = 1, b = 1$. Can you name this distribution?

(c) Determine and plot the density when $a = 2, b = 1$.

(d) Determine and plot the density when $a = 1, b = 2$.

(e) Determine and plot the density when $a = 2, b = 2$.

CHALLENGES

2.4.25 Prove (2.4.10). (Hint: Use $\Gamma(a)\Gamma(b) = \int_0^\infty \int_0^\infty x^{a-1}y^{b-1}e^{-x-y}\,dx\,dy$ and make the change of variable $u = x + y$, $v = x/u$.)

DISCUSSION TOPICS

2.4.26 Suppose $X \sim N(0, 1)$ and $Y \sim N(0, 4)$. Which do you think is larger, $P(X > 2)$ or $P(Y > 2)$? Why? (Hint: Look at Figure 2.4.5.)

2.5 Cumulative Distribution Functions

If X is a random variable, then its distribution consists of the values of $P(X \in B)$ for all subsets B of the real numbers. However, there are certain special subsets B that are convenient to work with. Specifically, if $B = (-\infty, x]$ for some real number x, then $P(X \in B) = P(X \le x)$. It turns out (see Theorem 2.5.1) that it is sufficient to keep track of $P(X \le x)$ for all real numbers x.

This motivates the following definition.

Definition 2.5.1 Given a random variable X, its *cumulative distribution function* (or *distribution function*, or *cdf* for short) is the function $F_X : R^1 \to [0, 1]$, defined by $F_X(x) = P(X \le x)$. (Where there is no confusion, we sometimes write $F(x)$ for $F_X(x)$.)

The reason for calling F_X the "distribution function" is that the full distribution of X can be determined directly from F_X. We demonstrate this for some events of particular importance.

First, suppose that $B = (a, b]$ is a left-open interval. Using (1.3.3),

$$P(X \in B) = P(a < X \le b) = P(X \le b) - P(X \le a) = F_X(b) - F_X(a).$$

Now, suppose that $B = [a, b]$ is a closed interval. Using the continuity of probability (see Theorem 1.6.1), we have

$$\begin{aligned} P(X \in B) &= P(a \le X \le b) = \lim_{n\to\infty} P(a - 1/n < X \le b) \\ &= \lim_{n\to\infty} (F_X(b) - F_X(a - 1/n)) = F_X(b) - \lim_{n\to\infty} F_X(a - 1/n). \end{aligned}$$

We sometimes write $\lim_{n\to\infty} F_X(a - 1/n)$ as $F_X(a^-)$, so that $P(X \in [a, b]) = F_X(b) - F_X(a^-)$. In the special case where $a = b$, we have

$$P(X = a) = F_X(a) - F_X(a^-). \tag{2.5.1}$$

Similarly, if $B = (a, b)$ is an open interval, then

$$P(X \in B) = P(a < X < b) = \lim_{n\to\infty} F_X(b - 1/n) - F_X(a) = F_X(b^-) - F_X(a).$$

If $B = [a, b)$ is a right-open interval, then

$$\begin{aligned} P(X \in B) &= P(a \le X < b) = \lim_{n\to\infty} F_X(b - 1/n) - \lim_{n\to\infty} F_X(a - 1/n) \\ &= F_X(b^-) - F_X(a^-). \end{aligned}$$

We conclude that we can determine $P(X \in B)$ from F_X whenever B is any kind of interval.

Now, if B is instead a *union* of intervals, then we can use additivity to again compute $P(X \in B)$ from F_X. For example, if

$$B = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_k, b_k],$$

with $a_1 < b_1 < a_2 < b_2 < \cdots < a_k < b_k$, then by additivity,

$$\begin{aligned} P(X \in B) &= P(X \in (a_1, b_1]) + \cdots + P(X \in (a_k, b_k]) \\ &= F_X(b_1) - F_X(a_1) + \cdots + F_X(b_k) - F_X(a_k). \end{aligned}$$

Hence, we can still compute $P(X \in B)$ solely from the values of $F_X(x)$.

Theorem 2.5.1 Let X be any random variable, with cumulative distribution function F_X. Let B be any subset of the real numbers. Then $P(X \in B)$ can be determined solely from the values of $F_X(x)$.

PROOF (Outline) It turns out that all relevant subsets B can be obtained by applying limiting operations to unions of intervals. Hence, because F_X determines $P(X \in B)$ when B is a union of intervals, it follows that F_X determines $P(X \in B)$ for all relevant subsets B. ▮

2.5.1 Properties of Distribution Functions

In light of Theorem 2.5.1, we see that cumulative distribution functions F_X are very useful. Thus, we note a few of their basic properties here.

Theorem 2.5.2 Let F_X be the cumulative distribution function of a random variable X. Then

(a) $0 \le F_X(x) \le 1$ for all x,
(b) $F_X(x) \le F_X(y)$ whenever $x \le y$ (i.e., F_X is increasing),
(c) $\lim_{x\to+\infty} F_X(x) = 1$,
(d) $\lim_{x\to-\infty} F_X(x) = 0$.

PROOF (a) Because $F_X(x) = P(X \le x)$ is a probability, it is between 0 and 1.
(b) Let $A = \{X \le x\}$ and $B = \{X \le y\}$. Then if $x \le y$, then $A \subseteq B$, so that

$P(A) \le P(B)$. But $P(A) = F_X(x)$ and $P(B) = F_X(y)$, so the result follows.

(c) Let $A_n = \{X \le n\}$. Because X must take on *some* value and hence $X \le n$ for sufficiently large n, we see that $\{A_n\}$ increases to S, i.e., $\{A_n\} \nearrow S$ (see Section 1.6). Hence, by continuity of P (see Theorem 1.6.1), $\lim_{n\to\infty} P(A_n) = P(S) = 1$. But $P(A_n) = P(X \le n) = F_X(n)$, so the result follows.

(d) Let $B_n = \{X \le -n\}$. Because $X \ge -n$ for sufficiently large n, $\{B_n\}$ decreases to the empty set, i.e., $\{B_n\} \searrow \emptyset$. Hence, again by continuity of P, $\lim_{n\to\infty} P(B_n) = P(\emptyset) = 0$. But $P(B_n) = P(X \le -n) = F_X(-n)$, so the result follows. ∎

If F_X is a cumulative distribution function, then F_X is also *right continuous*; see Problem 2.5.17. It turns out that if a function $F : R^1 \to R^1$ satisfies properties (a) through (d) and is right continuous, then there is a unique probability measure P on R^1 such that F is the cdf of P. We will not prove this result here.[2]

2.5.2 Cdfs of Discrete Distributions

We can compute the cumulative distribution function (cdf) F_X of a discrete random variable from its probability function p_X, as follows.

Theorem 2.5.3 Let X be a discrete random variable with probability function p_X. Then its cumulative distribution function F_X satisfies $F_X(x) = \sum_{y \le x} p_X(y)$.

PROOF Let $x_1, x_2, \ldots$ be the possible values of X. Then $F_X(x) = P(X \le x) = \sum_{x_i \le x} P(X = x_i) = \sum_{y \le x} P(X = y) = \sum_{y \le x} p_X(y)$, as claimed. ∎

Hence, if X is a discrete random variable, then by Theorem 2.5.3, F_X is piecewise constant, with a jump of size $p_X(x_i)$ at each value x_i. A plot of such a distribution looks like that depicted in Figure 2.5.1.

We consider an example of a distribution function of a discrete random variable.

EXAMPLE 2.5.1

Consider rolling one fair six-sided die, so that $S = \{1, 2, 3, 4, 5, 6\}$, with $P(s) = 1/6$ for each $s \in S$. Let X be the number showing on the die divided by 6, so that $X(s) = s/6$ for $s \in S$. What is $F_X(x)$? Since $X(s) \le x$ if and only if $s \le 6x$, we have that

$$F_X(x) = P(X \le x) = \sum_{s\in S,\, s\le 6x} P(s) = \sum_{s\in S,\, s\le 6x} \frac{1}{6} = \frac{1}{6} \left|\{s \in S : s \le 6x\}\right| .$$

[2] For example, see page 67 of *A First Look at Rigorous Probability Theory,* Second Edition, by J. S. Rosenthal (World Scientific Publishing, Singapore, 2006).

That is, to compute $F_X(x)$, we count how many elements $s \in S$ satisfy $s \le 6x$ and multiply that number by $1/6$. Therefore,

$$F_X(x) = \begin{cases} 0 & x < 1/6 \\ 1/6 & 1/6 \le x < 2/6 \\ 2/6 & 2/6 \le x < 3/6 \\ 3/6 & 3/6 \le x < 4/6 \\ 4/6 & 4/6 \le x < 5/6 \\ 5/6 & 5/6 \le x < 1 \\ 6/6 & 1 \le x. \end{cases}$$

In Figure 2.5.1, we present a graph of the function F_X and note that this is a step function. Note (see Exercise 2.5.1) that the properties of Theorem 2.5.2 are indeed satisfied by the function F_X.

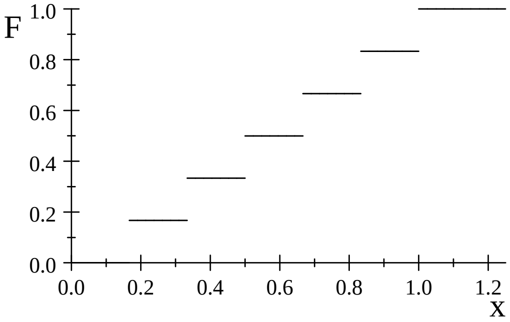

Figure 2.5.1: Graph of the cdf F_X in Example 2.5.1.

▮

2.5.3 Cdfs of Absolutely Continuous Distributions

Once we know the density f_X of X, then it is easy to compute the cumulative distribution function of X, as follows.

Theorem 2.5.4 Let X be an absolutely continuous random variable, with density function f_X. Then the cumulative distribution function F_X of X satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

for $x \in R^1$.

PROOF This follows from (2.4.5), by setting $b = x$ and letting $a \to -\infty$. ▮

From the fundamental theorem of calculus, we see that it is also possible to compute a density f_X once we know the cumulative distribution function F_X.

Corollary 2.5.1 Let X be an absolutely continuous random variable, with cumulative distribution function F_X. Let

$$f_X(x) = \frac{d}{dx} F_X(x) = F'_X(x).$$

Then f_X is a density function for X.

We note that F_X might not be differentiable everywhere, so that the function f_X of the corollary might not be defined at certain isolated points. The density function may take any value at such points.

Consider again the $N(0, 1)$ distribution, with density ϕ given by (2.4.9). According to Theorem 2.5.4, the cumulative distribution function F of this distribution is given by

$$F(x) = \int_{-\infty}^{x} \phi(t)\, dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

It turns out that it is provably impossible to evaluate this integral exactly, except for certain specific values of x (e.g., $x = -\infty$, $x = 0$, or $x = \infty$). Nevertheless, the cumulative distribution function of the $N(0, 1)$ distribution is so important that it is assigned a special symbol. Furthermore, this is tabulated in Table D.2 of Appendix D for certain values of x.

Definition 2.5.2 The symbol Φ stands for the cumulative distribution function of a standard normal distribution, defined by

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt, \tag{2.5.2}$$

for $x \in R^1$.

EXAMPLE 2.5.2 *Normal Probability Calculations*
Suppose that $X \sim N(0, 1)$, and we want to calculate

$$P(-0.63 \le X \le 2.0) = P(X \le 2.0) - P(X \le -0.63).$$

Then $P(X \le 2) = \Phi(2)$, while $P(X \le -0.63) = \Phi(-0.63)$. Unfortunately, $\Phi(2)$ and $\Phi(-0.63)$ cannot be computed exactly, but they can be approximated using a computer to numerically calculate the integral (2.5.2). Virtually all statistical software packages will provide such approximations, but many tabulations such as Table D.2, are also available. Using this table, we obtain $\Phi(2) = 0.9772$, while $\Phi(-0.63) = 0.2643$. This implies that

$$P(-0.63 \le X \le 2.0) = \Phi(2.0) - \Phi(-0.63) = 0.9772 - 0.2643 = 0.7129.$$

Now suppose that $X \sim N(\mu, \sigma^2)$, and we want to calculate $P(a \le X \le b)$. Letting f denote the density of X and following Example 2.4.8, we have

$$P(a \le X \le b) = \int_a^b f(x)\, dx = \int_a^b \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) dx.$$

Then, again following Example 2.4.8, we make the substitution $y = (x - \mu)/\sigma$ in the above integral to obtain

$$P(a \le X \le b) = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \phi(x)\,dx = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Therefore, general normal probabilities can be computed using the function Φ.

Suppose now that $a = -0.63$, $b = 2.0$, $\mu = 1.3$, and $\sigma^2 = 4$. We obtain

$$\begin{aligned} P(-0.63 \le X \le 2.0) &= \Phi\left(\frac{2.0 - 1.3}{2}\right) - \Phi\left(\frac{-0.63 - 1.3}{2}\right) \\ &= \Phi\,(0.35) - \Phi\,(-0.965) = 0.6368 - 0.16725 \\ &= 0.46955 \end{aligned}$$

because, using Table D.2, $\Phi\,(0.35) = 0.6368$. We approximate $\Phi\,(-0.965)$ by the linear interpolation between the values $\Phi\,(-0.96) = 0.1685$, $\Phi\,(-0.97) = 0.1660$, given by

$$\begin{aligned} \Phi\,(-0.965) &\approx \Phi\,(-0.96) + \frac{\Phi\,(-0.97) - \Phi\,(-0.96)}{-0.97 - (-0.96)}\,(-0.965 - (-0.96)) \\ &= 0.1685 + \frac{0.1660 - 0.1685}{-0.97 - (-0.96)}\,(-0.965 - (-0.96)) = 0.16725. \blacksquare \end{aligned}$$

EXAMPLE 2.5.3

Let X be a random variable with cumulative distribution function given by

$$F_X(x) = \begin{cases} 0 & x < 2 \\ (x-2)^4/16 & 2 \le x < 4 \\ 1 & 4 \le x. \end{cases}$$

In Figure 2.5.2, we present a graph of F_X.

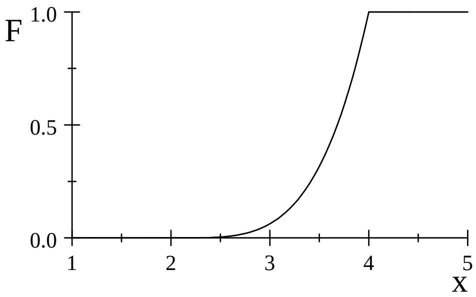

Figure 2.5.2: Graph of the cdf F_X in Example 2.5.3.

Suppose for this random variable X we want to compute $P(X \le 3)$, $P(X < 3)$, $P(X > 2.5)$, and $P(1.2 < X \le 3.4)$. We can compute all these probabilities directly from F_X. We have that

$$\begin{aligned} P(X \leq 3) &= F_X(3) = (3-2)^4/16 = 1/16, \\ P(X < 3) &= F_X(3^-) = \lim_{n\to\infty}(3-(1/n)-2)^4/16 = 1/16, \\ P(X > 2.5) &= 1 - P(X \leq 2.5) = 1 - F_X(2.5) \\ &= 1-(2.5-2)^4/16 = 1 - 0.0625/16 = 0.996, \\ P(1.2 < X \leq 3.4) &= F_X(3.4) - F_X(1.2) = (3.4-2)^4/16 - 0 = 0.2401. \blacksquare \end{aligned}$$

2.5.4 Mixture Distributions

Suppose now that $F_1, F_2, \ldots, F_k$ are cumulative distribution functions, corresponding to various distributions. Also let $p_1, p_2, \ldots, p_k$ be positive real numbers with $\sum_{i=1}^{k} p_i = 1$ (so these values form a probability distribution). Then we can define a new function G by

$$G(x) = p_1F_1(x) + p_2F_2(x) + \cdots + p_kF_k(x). \tag{2.5.3}$$

It is easily verified (see Exercise 2.5.6) that the function G given by (2.5.3) will satisfy properties (a) through (d) of Theorem 2.5.2 and is right continuous. Hence, G is also a cdf.

The distribution whose cdf is given by (2.5.3) is called a *mixture distribution* because it *mixes* the various distributions with cdfs $F_1, \ldots, F_k$ according to the probability distribution given by the $p_1, p_2, \ldots, p_k$.

To see how a mixture distribution arises in applications, consider a two-stage system, as discussed in Section 1.5.1. Let Z be a random variable describing the outcome of the first stage and such that $P(Z = i) = p_i$ for $i = 1, 2, \ldots, k$. Suppose that for the second stage, we observe a random variable Y where the distribution of Y depends on the outcome of the first stage, so that Y has cdf F_i when $Z = i$. In effect, F_i is the conditional distribution of Y, given that $Z = i$ (see Section 2.8). Then, by the law of total probability (see Theorem 1.5.1), the distribution function of Y is given by

$$P(Y \leq y) = \sum_{i=1}^{k} P(Y \leq y \mid Z = i)P(Z = i) = \sum_{i=1}^{k} p_i F_i\,(y) = G(y).$$

Therefore, the distribution function of Y is given by a mixture of the F_i.

Consider the following example of this.

EXAMPLE 2.5.4

Suppose we have two bowls containing chips. Bowl #1 contains one chip labelled 0, two chips labelled 3, and one chip labelled 5. Bowl #2 contains one chip labelled 2, one chip labelled 4, and one chip labelled 5. Now let X_i be the random variable corresponding to randomly drawing a chip from bowl #i. Therefore, $P(X_1 = 0) = 1/4$, $P(X_1 = 3) = 1/2$, and $P(X_1 = 5) = 1/4$, while $P(X_2 = 2) = P(X_2 = 4) =$

$P(X_2 = 5) = 1/3$. Then X_1 has distribution function given by

$$F_1(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \le x < 3 \\ 3/4 & 3 \le x < 5 \\ 1 & x \ge 5 \end{cases}$$

and X_2 has distribution function given by

$$F_2(x) = \begin{cases} 0 & x < 2 \\ 1/3 & 2 \le x < 4 \\ 2/3 & 4 \le x < 5 \\ 1 & x \ge 5. \end{cases}$$

Now suppose that we choose a bowl by randomly selecting a card from a deck of five cards where one card is labelled 1 and four cards are labelled 2. Let Z denote the value on the card obtained, so that $P(Z = 1) = 1/5$ and $P(Z = 2) = 4/5$. Then, having obtained the value $Z = i$, we observe Y by randomly drawing a chip from bowl #i. We see immediately that the cdf of Y is given by

$$G(x) = (1/5)\, F_1(x) + (4/5)\, F_2(x),$$

and this is a mixture of the cdfs F_1 and F_2. ▮

As the following examples illustrate, it is also possible to have *infinite* mixtures of distributions.

EXAMPLE 2.5.5 *Location and Scale Mixtures*

Suppose F is some cumulative distribution function. Then for any real number y, the function F_y defined by $F_y(x) = F(x - y)$ is also a cumulative distribution function. In fact, F_y is just a "shifted" version of F. An example of this is depicted in Figure 2.5.3.

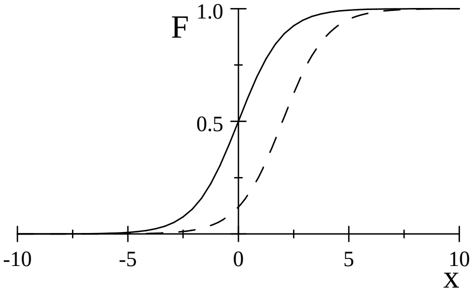

Figure 2.5.3: Plot of the distribution functions F (solid line) and F_2 (dashed line) in Example 2.5.5, where $F(x) = e^x / (e^x + 1)$ for $x \in R^1$.

If $p_i \ge 0$ with $\sum_i p_i = 1$ (so the p_i form a probability distribution), and $y_1, y_2, \ldots$ are real numbers, then we can define a *discrete location mixture* by

$$H(x) = \sum_i p_i F_{y_i}(x) = \sum_i p_i F(x - y_i).$$

Indeed, the shift $F_y(x) = F(x - y)$ itself corresponds to a special case of a discrete location mixture, with $p_1 = 1$ and $y_1 = y$.

Furthermore, if g is some nonnegative function with $\int_{-\infty}^{\infty} g(t)\, dt = 1$ (so g is a density function), then we can define

$$H(x) = \int_{-\infty}^{\infty} F_y(x)\, g(y)\, dy = \int_{-\infty}^{\infty} F(x - y)\, g(y)\, dy.$$

Then it is not hard to see that H is also a cumulative distribution function — one that is called a *continuous location mixture* of F. The idea is that H corresponds to a *mixture* of different shifted distributions F_y, with the density g giving the distribution of the mixing coefficient y.

We can also define a *discrete scale mixture* by

$$K(x) = \sum_i p_i F(x/y_i)$$

whenever $y_i > 0$, $p_i \geq 0$, and $\sum_i p_i = 1$. Similarly, if $\int_0^{\infty} g(t)\, dt = 1$, then we can write

$$K(x) = \int_0^{\infty} F(x/y) g(y)\, dy.$$

Then K is also a cumulative distribution function, called a *continuous scale mixture* of F. ▮

You might wonder at this point whether a mixture distribution is discrete or continuous. The answer depends on the distributions being mixed and the mixing distribution. For example, discrete location mixtures of discrete distributions are discrete and discrete location mixtures of continuous distributions are continuous.

There is nothing restricting us, however, to mixing only discrete distributions or only continuous distributions. Other kinds of distribution are considered in the following section.

2.5.5 Distributions Neither Discrete Nor Continuous (Advanced)

There are some distributions that are neither discrete nor continuous, as the following example shows.

EXAMPLE 2.5.6

Suppose that $X_1 \sim$ Poisson(3) is discrete with cdf F_1, while $X_2 \sim N(0, 1)$ is continuous with cdf F_2, and Y has the mixture distribution given by $F_Y(y) = (1/5)\, F_1(y) + (4/5)\, F_2(y)$. Using (2.5.1), we have

$$\begin{aligned} P(Y = y) &= F_Y(y) - F_Y(y^-) \\ &= (1/5)F_1(y) + (4/5)\, F_2(y) - (1/5)F_1(y^-) - (4/5)F_2(y^-) \\ &= (1/5)\left(F_1(y) - F_1(y^-)\right) + (4/5)\left(F_2(y) - F_2(y^-)\right) \\ &= \frac{1}{5}P(X_1 = y) + \frac{4}{5}P(X_2 = y). \end{aligned}$$

Therefore,

$$P(Y = y) = \begin{cases} \frac{1}{5}\frac{3^y}{y!}e^{-3} & y \text{ a nonnegative integer} \\ 0 & \text{otherwise.} \end{cases}$$

Because $P(Y = y) > 0$ for nonnegative integers y, the random variable Y is not continuous. On the other hand, we have

$$\sum_y P(Y = y) = \sum_{y=0}^{\infty} \frac{1}{5}\frac{3^y}{y!}e^{-3} = \frac{1}{5} < 1.$$

Hence, Y is not discrete either.

In fact, Y is neither discrete nor continuous. Rather, Y is a mixture of a discrete and a continuous distribution. ▮

For the most part in this book, we shall treat discrete and continuous distributions separately. However, it is important to keep in mind that actual distributions may be neither discrete nor continuous but rather a mixture of the two.[3] In most applications, however, the distributions we deal with are either continuous or discrete.

Recall that a continuous distribution need not be absolutely continuous, i.e., have a density. Hence, a distribution that is a mixture of a discrete and a continuous distribution might not be a mixture of a discrete and an absolutely continuous distribution.

Summary of Section 2.5

- The cumulative distribution function (cdf) of X is $F_X(x) = P(X \le x)$.
- All probabilities associated with X can be determined from F_X.
- As x increases from $-\infty$ to ∞, $F_X(x)$ increases from 0 to 1.
- If X is discrete, then $F_X(x) = \sum_{y \le x} P(X = y)$.
- If X is absolutely continuous, then $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$, and $f_X(x) = F_X'(x)$.
- We write $\Phi(x)$ for the cdf of the standard normal distribution evaluated at x.
- A mixture distribution has a cdf that is a linear combination of other cdfs. Two special cases are location and scale mixtures.
- Some mixture distributions are neither discrete nor continuous.

EXERCISES

2.5.1 Verify explicitly that properties (a) through (d) of Theorem 2.5.2 are indeed satisfied by the function F_X in Example 2.5.1.

2.5.2 Consider rolling one fair six-sided die, so that $S = \{1, 2, 3, 4, 5, 6\}$, and $P(s) = 1/6$ for all $s \in S$. Let X be the number showing on the die, so that $X(s) = s$ for $s \in S$. Let $Y = X^2$. Compute the cumulative distribution function $F_Y(y) = P(Y \le y)$, for all $y \in R^1$. Verify explicitly that properties (a) through (d) of Theorem 2.5.2 are satisfied by this function F_Y.

[3] In fact, there exist probability distributions that cannot be expressed even as a mixture of a discrete and a continuous distribution, but these need not concern us here.

2.5.3 For each of the following functions F, determine whether or not F is a valid cumulative distribution function, i.e., whether or not F satisfies properties (a) through (d) of Theorem 2.5.2.

(a) $F(x) = x$ for all $x \in R^1$

(b)

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

(c)

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

(d)

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

(e)

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2/9 & 0 \le x \le 3 \\ 1 & x > 3 \end{cases}$$

(f)

$$F(x) = \begin{cases} 0 & x < 1 \\ x^2/9 & 1 \le x \le 3 \\ 1 & x > 3 \end{cases}$$

(g)

$$F(x) = \begin{cases} 0 & x < -1 \\ x^2/9 & -1 \le x \le 3 \\ 1 & x > 3 \end{cases}$$

2.5.4 Let $X \sim N(0, 1)$. Compute each of the following in terms of the function Φ of Definition 2.5.2 and use Table D.2 (or software) to evaluate these probabilities numerically.

(a) $P(X \le -5)$

(b) $P(-2 \le X \le 7)$

(c) $P(X \ge 3)$

2.5.5 Let $Y \sim N(-8, 4)$. Compute each of the following, in terms of the function Φ of Definition 2.5.2 and use Table D.2 (or software) to evaluate these probabilities numerically.

(a) $P(Y \le -5)$

(b) $P(-2 \le Y \le 7)$

(c) $P(Y \ge 3)$

2.5.6 Verify that the function G given by (2.5.3) satisfies properties (a) through (d) of Theorem 2.5.2.

2.5.7 Suppose $F_X(x) = x^2$ for $0 \le x \le 1$. Compute each of the following.
(a) $P(X < 1/3)$
(b) $P(1/4 < X < 1/2)$
(c) $P(2/5 < X < 4/5)$
(d) $P(X < 0)$
(e) $P(X < 1)$
(f) $P(X < -1)$
(g) $P(X < 3)$
(h) $P(X = 3/7)$

2.5.8 Suppose $F_Y(y) = y^3$ for $0 \le y < 1/2$, and $F_Y(y) = 1 - (1 - y)^3$ for $1/2 \le y \le 1$. Compute each of the following.
(a) $P(1/3 < Y < 3/4)$
(b) $P(Y = 1/3)$
(c) $P(Y = 1/2)$

2.5.9 Let $F(x) = x^2$ for $0 \le x \le 2$, with $F(x) = 0$ for $x < 0$ and $F(x) = 4$ for $x > 2$.
(a) Sketch a graph of F.
(b) Is F a valid cumulative distribution function? Why or why not?

2.5.10 Let $F(x) = 0$ for $x < 0$, with $F(x) = e^{-x}$ for $x \ge 0$.
(a) Sketch a graph of F.
(b) Is F a valid cumulative distribution function? Why or why not?

2.5.11 Let $F(x) = 0$ for $x < 0$, with $F(x) = 1 - e^{-x}$ for $x \ge 0$.
(a) Sketch a graph of F.
(b) Is F a valid cumulative distribution function? Why or why not?

2.5.12 Let $X \sim$ Exponential(3). Compute the function F_X.

2.5.13 Let $F(x) = 0$ for $x < 0$, with $F(x) = 1/3$ for $0 \le x < 2/5$, and $F(x) = 3/4$ for $2/5 \le x < 4/5$, and $F(x) = 1$ for $x \ge 4/5$.
(a) Sketch a graph of F.
(b) Prove that F is a valid cumulative distribution function.
(c) If X has cumulative distribution function equal to F, then compute $P(X > 4/5)$ and $P(-1 < X < 1/2)$ and $P(X = 2/5)$ and $P(X = 4/5)$.

2.5.14 Let $G(x) = 0$ for $x < 0$, with $G(x) = 1 - e^{-x^2}$ for $x \ge 0$.
(a) Prove that G is a valid cumulative distribution function.
(b) If Y has cumulative distribution function equal to G, then compute $P(Y > 4)$ and $P(-1 < Y < 2)$ and $P(Y = 0)$.

2.5.15 Let F and G be as in the previous two exercises. Let $H(x) = (1/3)F(x) + (2/3)G(x)$. Suppose Z has cumulative distribution function equal to H. Compute each of the following.
(a) $P(Z > 4/5)$
(b) $P(-1 < Z < 1/2)$
(c) $P(Z = 2/5)$
(d) $P(Z = 4/5)$

(e) $P(Z = 0)$
(f) $P(Z = 1/2)$

PROBLEMS

2.5.16 Let F be a cumulative distribution function. Compute (with explanation) the value of $\lim_{n\to\infty}[F(2n) - F(n)]$.

2.5.17 Let F be a cumulative distribution function. For $x \in R^1$, we could define $F(x^+)$ by $F(x^+) = \lim_{n\to\infty} F(x + \frac{1}{n})$. Prove that F is *right continuous*, meaning that for each $x \in R^1$, we have $F(x^+) = F(x)$. (Hint: You will need to use continuity of P (Theorem 1.6.1).)

2.5.18 Let X be a random variable, with cumulative distribution function F_X. Prove that $P(X = a) = 0$ if and only if the function F_X is continuous at a. (Hint: Use (2.5.1) and the previous problem.)

2.5.19 Let Φ be as in Definition 2.5.2. Derive a formula for $\Phi(-x)$ in terms of $\Phi(x)$. (Hint: Let $s = -t$ in (2.5.2), and do not forget Theorem 2.5.2.)

2.5.20 Determine the distribution function for the logistic distribution of Problem 2.4.18.

2.5.21 Determine the distribution function for the Weibull(α) distribution of Problem 2.4.19.

2.5.22 Determine the distribution function for the Pareto(α) distribution of Problem 2.4.20.

2.5.23 Determine the distribution function for the Cauchy distribution of Problem 2.4.21.

2.5.24 Determine the distribution function for the Laplace distribution of Problem 2.4.22.

2.5.25 Determine the distribution function for the extreme value distribution of Problem 2.4.23.

2.5.26 Determine the distribution function for the beta distributions of Problem 2.4.24 for parts (b) through (e).

DISCUSSION TOPICS

2.5.27 Does it surprise you that all information about the distribution of a random variable X can be stored by a single function F_X? Why or why not? What other examples can you think of where lots of different information is stored by a single function?

2.6 One-Dimensional Change of Variable

Let X be a random variable with a known distribution. Suppose that $Y = h(X)$, where $h : R^1 \to R^1$ is some function. (Recall that this really means that $Y(s) = h(X(s))$, for all $s \in S$.) Then what is the distribution of Y?

2.6.1 The Discrete Case

If X is a *discrete* random variable, this is quite straightforward. To compute the probability that $Y = y$, we need to compute the probability of the set consisting of all the x values satisfying $h(x) = y$, namely, compute $P\,(X \in \{x : h(x) = y\})$. This is depicted graphically in Figure 2.6.1.

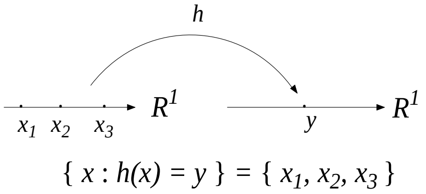

Figure 2.6.1: An example where the set of x values that satisfy $h(x) = y$ consists of three points x_1, x_2, and x_3.

We now establish the basic result.

Theorem 2.6.1 Let X be a discrete random variable, with probability function p_X. Let $Y = h(X)$, where $h : R^1 \to R^1$ is some function. Then Y is also discrete, and its probability function p_Y satisfies $p_Y(y) = \sum_{x \in h^{-1}\{y\}} p_X(x)$, where $h^{-1}\{y\}$ is the set of all real numbers x with $h(x) = y$.

PROOF We compute that $p_Y(y) = P(h(X) = y) = \sum_{x \in h^{-1}\{y\}} P(X = x) = \sum_{x \in h^{-1}\{y\}} p_X(x)$, as claimed. ▮

EXAMPLE 2.6.1

Let X be the number of heads when flipping three fair coins. Let $Y = 1$ if $X \geq 1$, with $Y = 0$ if $X = 0$. Then $Y = h(X)$, where $h(0) = 0$ and $h(1) = h(2) = h(3) = 1$. Hence, $h^{-1}\{0\} = \{0\}$, so $P(Y = 0) = P(X = 0) = 1/8$. On the other hand, $h^{-1}\{1\} = \{1, 2, 3\}$, so $P(Y = 1) = P(X = 1) + P(X = 2) + P(X = 3) = 3/8 + 3/8 + 1/8 = 7/8$. ▮

EXAMPLE 2.6.2

Let X be the number showing on a fair six-sided die, so that $P(X = x) = 1/6$ for $x = 1, 2, 3, 4, 5$, and 6. Let $Y = X^2 - 3X + 2$. Then $Y = h(X)$, where $h(x) = x^2 - 3x + 2$. Note that $h(x) = 0$ if and only if $x = 1$ or $x = 2$. Hence, $h^{-1}\{0\} = \{1, 2\}$ and

$$P(Y = 0) = p_X(1) + p_X(2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}. \text{ ▮}$$

2.6.2 The Continuous Case

If X is *continuous* and $Y = h(X)$, then the situation is more complicated. Indeed, Y might not be continuous at all, as the following example shows.

EXAMPLE 2.6.3

Let X have the uniform distribution on $[0, 1]$, i.e., $X \sim \text{Uniform}[0, 1]$, as in Example 2.4.2. Let $Y = h(X)$, where

$$h(x) = \begin{cases} 7 & x \leq 3/4 \\ 5 & x > 3/4. \end{cases}$$

Here, $Y = 7$ if and only if $X \leq 3/4$ (which happens with probability 3/4), whereas $Y = 5$ if and only if $X > 3/4$ (which happens with probability 1/4). Hence, Y is discrete, with probability function p_Y satisfying $p_Y(7) = 3/4$, $p_Y(5) = 1/4$, and $p_Y(y) = 0$ when $y \neq 5, 7$. ▮

On the other hand, if X is absolutely continuous, and the function h is *strictly increasing*, then the situation is considerably simpler, as the following theorem shows.

Theorem 2.6.2 Let X be an absolutely continuous random variable, with density function f_X. Let $Y = h(X)$, where $h : R^1 \to R^1$ is a function that is differentiable and strictly increasing. Then Y is also absolutely continuous, and its density function f_Y is given by

$$f_Y(y) = f_X(h^{-1}(y)) \,/\, |h'(h^{-1}(y))|, \tag{2.6.1}$$

where h' is the derivative of h, and where $h^{-1}(y)$ is the unique number x such that $h(x) = y$.

PROOF See Section 2.11 for the proof of this result. ▮

EXAMPLE 2.6.4

Let $X \sim \text{Uniform}[0, 1]$, and let $Y = 3X$. What is the distribution of Y?

Here, X has density f_X, given by $f_X(x) = 1$ if $0 \leq x \leq 1$, and $f_X(x) = 0$ otherwise. Also, $Y = h(X)$, where h is defined by $h(x) = 3x$. Note that h is strictly increasing because if $x < y$, then $3x < 3y$, i.e., $h(x) < h(y)$. Hence, we may apply Theorem 2.6.2.

We note first that $h'(x) = 3$ and that $h^{-1}(y) = y/3$. Then, according to Theorem 2.6.2, Y is absolutely continuous with density

$$\begin{aligned} f_Y(y) &= f_X(h^{-1}(y))/|h'(h^{-1}(y))| = \frac{1}{3} f_X(y/3) \\ &= \begin{cases} 1/3 & 0 \leq y/3 \leq 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1/3 & 0 \leq y \leq 3 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

By comparison with Example 2.4.3, we see that $Y \sim \text{Uniform}[0, 3]$, i.e., that Y has the $\text{Uniform}[L, R]$ distribution with $L = 0$ and $R = 3$. ▮

EXAMPLE 2.6.5

Let $X \sim N(0, 1)$, and let $Y = 2X + 5$. What is the distribution of Y?

Here, X has density f_X, given by

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Also, $Y = h(X)$, where h is defined by $h(x) = 2x + 5$. Note that again, h is strictly increasing because if $x < y$, then $2x + 5 < 2y + 5$, i.e., $h(x) < h(y)$. Hence, we may again apply Theorem 2.6.2.

We note first that $h'(x) = 2$ and that $h^{-1}(y) = (y - 5)/2$. Then, according to Theorem 2.6.2, Y is absolutely continuous with density

$$f_Y(y) = f_X(h^{-1}(y))/|h'(h^{-1}(y))| = f_X((y-5)/2)/2 = \frac{1}{2\sqrt{2\pi}} e^{-(y-5)^2/8}.$$

By comparison with Example 2.4.8, we see that $Y \sim N(5, 4)$, i.e., that Y has the $N(\mu, \sigma^2)$ distribution with $\mu = 5$ and $\sigma^2 = 4$. ▮

If instead the function h is strictly *decreasing*, then a similar result holds.

Theorem 2.6.3 Let X be an absolutely continuous random variable, with density function f_X. Let $Y = h(X)$, where $h : R^1 \to R^1$ is a function that is differentiable and strictly decreasing. Then Y is also absolutely continuous, and its density function f_Y may again be defined by (2.6.1).

PROOF See Section 2.11 for the proof of this result. ▮

EXAMPLE 2.6.6

Let $X \sim$ Uniform[0, 1], and let $Y = \ln(1/X)$. What is the distribution of Y?

Here, X has density f_X, given by $f_X(x) = 1$ for $0 \le x \le 1$, and $f_X(x) = 0$ otherwise. Also, $Y = h(X)$, where h is defined by $h(x) = \ln(1/x)$. Note that here, h is strictly decreasing because if $x < y$, then $1/x > 1/y$, so $\ln(1/x) > \ln(1/y)$, i.e., $h(x) > h(y)$. Hence, we may apply Theorem 2.6.3.

We note first that $h'(x) = -1/x$ and that $h^{-1}(y) = e^{-y}$. Then, by Theorem 2.6.3, Y is absolutely continuous with density

$$\begin{aligned} f_Y(y) &= f_X(h^{-1}(y))/|h'(h^{-1}(y))| = e^{-y}\, f_X(e^{-y}) \\ &= \begin{cases} e^{-y} & 0 \le e^{-y} \le 1 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} e^{-y} & y \ge 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

By comparison with Example 2.4.4, we see that $Y \sim$ Exponential(1), i.e., that Y has the Exponential(1) distribution. ▮

Finally, we note the following.

Theorem 2.6.4 Theorem 2.6.2 (and 2.6.3) remains true assuming only that h is strictly increasing (or decreasing) at places for which $f_X(x) > 0$. If $f_X(x) = 0$ for an interval of x values, then it does not matter how the function h behaves in that interval (or even if it is well defined there).

EXAMPLE 2.6.7

If $X \sim$ Exponential(λ), then $f_X(x) = 0$ for $x < 0$. Therefore, it is required that h be strictly increasing (or decreasing) only for $x \ge 0$. Thus, functions such as $h(x) = x^2$, $h(x) = x^8$, and $h(x) = \sqrt{x}$ could still be used with Theorem 2.6.2, while functions such as $h(x) = -x^2$, $h(x) = -x^8$, and $h(x) = -\sqrt{x}$ could still be used with Theorem 2.6.3, even though such functions may not necessarily be strictly increasing (or decreasing) and well defined on the entire real line. ▮

Summary of Section 2.6

- If X is discrete, and $Y = h(X)$, then $P(Y = y) = \sum_{x:\, h(x)=y} P(X = x)$.
- If X is absolutely continuous, and $Y = h(X)$ with h strictly increasing or strictly decreasing, then the density of Y is given by $f_Y(y) = f_X(h^{-1}(y)) \,/\, |h'(h^{-1}(y))|$.
- This allows us to compute the distribution of a function of a random variable.

EXERCISES

2.6.1 Let $X \sim \text{Uniform}[L, R]$. Let $Y = cX + d$, where $c > 0$. Prove that $Y \sim \text{Uniform}[cL + d, cR + d]$. (This generalizes Example 2.6.4.)

2.6.2 Let $X \sim \text{Uniform}[L, R]$. Let $Y = cX + d$, where $c < 0$. Prove that $Y \sim \text{Uniform}[cR + d, cL + d]$. (In particular, if $L = 0$ and $R = 1$ and $c = -1$ and $d = 1$, then $X \sim \text{Uniform}[0, 1]$ and also $Y = 1 - X \sim \text{Uniform}[0, 1]$.)

2.6.3 Let $X \sim N(\mu, \sigma^2)$. Let $Y = cX + d$, where $c > 0$. Prove that $Y \sim N(c\mu + d, c^2\sigma^2)$. (This generalizes Example 2.6.5.)

2.6.4 Let $X \sim \text{Exponential}(\lambda)$. Let $Y = cX$, where $c > 0$. Prove that $Y \sim \text{Exponential}(\lambda/c)$.

2.6.5 Let $X \sim \text{Exponential}(\lambda)$. Let $Y = X^3$. Compute the density f_Y of Y.

2.6.6 Let $X \sim \text{Exponential}(\lambda)$. Let $Y = X^{1/4}$. Compute the density f_Y of Y. (Hint: Use Theorem 2.6.4.)

2.6.7 Let $X \sim \text{Uniform}[0, 3]$. Let $Y = X^2$. Compute the density function f_Y of Y.

2.6.8 Let X have a density such that $f_X(\mu + x) = f_X(\mu - x)$, i.e., it is symmetric about μ. Let $Y = 2\mu - X$. Show that the density of Y is given by f_X. Use this to determine the distribution of Y when $X \sim N(\mu, \sigma^2)$.

2.6.9 Let X have density function $f_X(x) = x^3/4$ for $0 < x < 2$, otherwise $f_X(x) = 0$.
(a) Let $Y = X^2$. Compute the density function $f_Y(y)$ for Y.
(b) Let $Z = \sqrt{X}$. Compute the density function $f_Z(z)$ for Z.

2.6.10 Let $X \sim \text{Uniform}[0, \pi/2]$. Let $Y = \sin(X)$. Compute the density function $f_Y(y)$ for Y.

2.6.11 Let X have density function $f_X(x) = (1/2)\sin(x)$ for $0 < x < \pi$, otherwise $f_X(x) = 0$. Let $Y = X^2$. Compute the density function $f_Y(y)$ for Y.

2.6.12 Let X have density function $f_X(x) = 1/x^2$ for $x > 1$, otherwise $f_X(x) = 0$. Let $Y = X^{1/3}$. Compute the density function $f_Y(y)$ for Y.

2.6.13 Let $X \sim \text{Normal}(0, 1)$. Let $Y = X^3$. Compute the density function $f_Y(y)$ for Y.

PROBLEMS

2.6.14 Let $X \sim \text{Uniform}[2, 7]$, $Y = X^3$, and $Z = \sqrt{Y}$. Compute the density f_Z of Z, in two ways.
(a) Apply Theorem 2.6.2 to first obtain the density of Y, then apply Theorem 2.6.2 *again* to obtain the density of Z.
(b) Observe that $Z = \sqrt{Y} = \sqrt{X^3} = X^{3/2}$, and apply Theorem 2.6.2 just *once*.

2.6.15 Let $X \sim \text{Uniform}[L, R]$, and let $Y = h(X)$ where $h(x) = (x - c)^6$. According to Theorem 2.6.4, under what conditions on L, R, and c can we apply Theorem 2.6.2 or Theorem 2.6.3 to this choice of X and Y?

2.6.16 Let $X \sim N(\mu, \sigma^2)$. Let $Y = cX + d$, where $c < 0$. Prove that again $Y \sim N(c\mu + d, c^2\sigma^2)$, just like in Exercise 2.6.3.

2.6.17 (*Log-normal*(τ) *distribution*) Suppose that $X \sim N(0, \tau^2)$. Prove that $Y = e^X$ has density

$$f_\tau(y) = \frac{1}{\sqrt{2\pi}\,\tau} \exp\left(-\frac{(\ln y)^2}{2\tau^2}\right)\frac{1}{y}$$

for $y > 0$ and where $\tau > 0$ is unknown. We say that $Y \sim \text{Log-normal}(\tau)$.

2.6.18 Suppose that $X \sim \text{Weibull}(\alpha)$ (see Problem 2.4.19). Determine the distribution of $Y = X^\beta$.

2.6.19 Suppose that $X \sim \text{Pareto}(\alpha)$ (see Problem 2.4.20). Determine the distribution of $Y = (1 + X)^\beta - 1$.

2.6.20 Suppose that X has the extreme value distribution (see Problem 2.4.23). Determine the distribution of $Y = e^{-X}$.

CHALLENGES

2.6.21 Theorems 2.6.2 and 2.6.3 require that h be an increasing or decreasing function, at least at places where the density of X is positive (see Theorem 2.6.4). Suppose now that $X \sim N(0, 1)$ and $Y = h(X)$, where $h(x) = x^2$. Then $f_X(x) > 0$ for all x, while h is increasing only for $x > 0$ and decreasing only for $x < 0$. Hence, Theorems 2.6.2 and 2.6.3 do not directly apply. Compute $f_Y(y)$ anyway. (Hint: $P(a \le Y \le b) = P(a \le Y \le b,\ X > 0) + P(a \le Y \le b,\ X < 0)$.)

2.7 Joint Distributions

Suppose X and Y are two random variables. Even if we know the distributions of X and Y exactly, this still does not tell us anything about the *relationship* between X and Y.

EXAMPLE 2.7.1

Let $X \sim \text{Bernoulli}(1/2)$, so that $P(X = 0) = P(X = 1) = 1/2$. Let $Y_1 = X$, and let $Y_2 = 1 - X$. Then we clearly have $Y_1 \sim \text{Bernoulli}(1/2)$ and $Y_2 \sim \text{Bernoulli}(1/2)$ as well.

On the other hand, the *relationship* between X and Y_1 is very different from the relationship between X and Y_2. For example, if we know that $X = 1$, then we also must have $Y_1 = 1$, but $Y_2 = 0$. Hence, merely knowing that X, Y_1, and Y_2 all have the distribution Bernoulli$(1/2)$ does not give us complete information about the relationships among these random variables. ▮

A formal definition of joint distribution is as follows.

Definition 2.7.1 If X and Y are random variables, then the *joint distribution* of X and Y is the collection of probabilities $P((X, Y) \in B)$, for all subsets $B \subseteq R^2$ of pairs of real numbers.

Joint distributions, like other distributions, are so complicated that we use various functions to describe them, including joint cumulative distribution functions, joint probability functions, and joint density functions, as we now discuss.

2.7.1 Joint Cumulative Distribution Functions

Definition 2.7.2 Let X and Y be random variables. Then their *joint cumulative distribution function* is the function $F_{X,Y} : R^2 \to [0, 1]$ defined by

$$F_{X,Y}(x, y) = P(X \le x,\ Y \le y).$$

(Recall that the comma means "and" here, so that $F_{X,Y}(x, y)$ is the probability that $X \le x$ *and* $Y \le y$.)

EXAMPLE 2.7.2 (*Example 2.7.1 continued*)
Again, let $X \sim$ Bernoulli$(1/2)$, $Y_1 = X$, and $Y_2 = 1 - X$. Then we compute that

$$F_{X,Y_1}(x, y) = P(X \le x, Y_1 \le y) = \begin{cases} 0 & \min(x, y) < 0 \\ 1/2 & 0 \le \min(x, y) < 1 \\ 1 & \min(x, y) \ge 1. \end{cases}$$

On the other hand,

$$F_{X,Y_2}(x, y) = P(X \le x, Y_2 \le y) = \begin{cases} 0 & \min(x, y) < 0 \text{ or } \max(x, y) < 1 \\ 1/2 & 0 \le \min(x, y) < 1 \le \max(x, y) \\ 1 & \min(x, y) \ge 1. \end{cases}$$

We thus see that F_{X,Y_1} is quite a different function from F_{X,Y_2}. This reflects the fact that, even though Y_1 and Y_2 each have the same distribution, their relationship with X is quite different. On the other hand, the functions F_{X,Y_1} and F_{X,Y_2} are rather cumbersome and awkward to work with. ▮

We see from this example that joint cumulative distribution functions (or joint cdfs) do indeed keep track of the relationship between X and Y. Indeed, joint cdfs tell us everything about the joint probabilities of X and Y, as the following theorem (an analog of Theorem 2.5.1) shows.

Theorem 2.7.1 Let X and Y be any random variables, with joint cumulative distribution function $F_{X,Y}$. Let B be a subset of R^2. Then $P((X, Y) \in B)$ can be determined solely from the values of $F_{X,Y}(x, y)$.

We shall not give a proof of Theorem 2.7.1, although it is similar to the proof of Theorem 2.5.1. However, the following theorem indicates why Theorem 2.7.1 is true, and it also provides a useful computational fact.

Theorem 2.7.2 Let X and Y be any random variables, with joint cumulative distribution function $F_{X,Y}$. Suppose $a \le b$ and $c \le d$. Then

$$P(a < X \le b\,,\ c < Y \le d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c).$$

PROOF According to (1.3.3),

$$\begin{aligned}&P(a < X \le b,\ c < Y \le d)\\ &= P(X \le b,\ Y \le d) - P(X \le b,\ Y \le d\text{ , and either } X \le a \text{ or } Y \le c).\end{aligned}$$

But by the principle of inclusion–exclusion (1.3.4),

$$\begin{aligned}&P(X \le b,\ Y \le d,\ \text{ and either } X \le a \text{ or } Y \le c)\\ &= P(X \le b,\ Y \le c) + P(X \le a,\ Y \le d) - P(X \le a,\ Y \le c).\end{aligned}$$

Combining these two equations, we see that

$$\begin{aligned}&P(a < X \le b, c < Y \le d)\\ &= P(X \le b, Y \le d) - P(X \le a, Y \le d) - P(X \le b, Y \le c) + P(X \le a, Y \le c)\end{aligned}$$

and from this we obtain

$$P(a < X \le b, c < Y \le d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c),$$

as claimed. ∎

Joint cdfs are not easy to work with. Thus, in this section we shall also consider other functions, which are more convenient for pairs of discrete or absolutely continuous random variables.

2.7.2 Marginal Distributions

We have seen how a joint cumulative distribution function $F_{X,Y}$ tells us about the relationship between X and Y. However, the function $F_{X,Y}$ also tells us everything about each of X and Y separately, as the following theorem shows.

Theorem 2.7.3 Let X and Y be two random variables, with joint cumulative distribution function $F_{X,Y}$. Then the cumulative distribution function F_X of X satisfies

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y),$$

for all $x \in R^1$. Similarly, the cumulative distribution function F_Y of Y satisfies

$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y),$$

for all $y \in R^1$.

PROOF Note that we *always* have $Y \le \infty$. Hence, using continuity of P, we have

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P(X \le x,\ Y \le \infty) \\ &= \lim_{y \to \infty} P(X \le x,\ Y \le y) \\ &= \lim_{y \to \infty} F_{X,Y}(x, y), \end{aligned}$$

as claimed. Similarly,

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(X \le \infty,\ Y \le y) \\ &= \lim_{x \to \infty} P(X \le x,\ Y \le y) \\ &= \lim_{x \to \infty} F_{X,Y}(x, y), \end{aligned}$$

completing the proof. ▮

In the context of Theorem 2.7.3, F_X is called the *marginal cumulative distribution function* of X, and the distribution of X is called the *marginal distribution* of X. (Similarly, F_Y is called the marginal cumulative distribution function of Y, and the distribution of Y is called the marginal distribution of Y.) Intuitively, if we think of $F_{X,Y}$ as being a function of a pair (x, y), then F_X and F_Y are functions of x and y, respectively, which could be written into the "margins" of a graph of $F_{X,Y}$.

EXAMPLE 2.7.3

In Figure 2.7.1, we have plotted the joint distribution function

$$F_{X,Y}(x, y) = \begin{cases} 0 & x < 0 \text{ or } y < 0 \\ xy^2 & 0 \le x \le 1, 0 \le y \le 1 \\ x & 0 \le x \le 1, y \ge 1 \\ y^2 & x \ge 1, 0 \le y \le 1 \\ 1 & x > 1 \text{ and } y > 1. \end{cases}$$

It is easy to see that

$$F_X(x) = F_{X,Y}(x, 1) = x$$

for $0 \le x \le 1$ and that

$$F_Y(y) = F_{X,Y}(1, y) = y^2$$

for $0 \le y \le 1$. The graphs of these functions are given by the outermost edges of the surface depicted in Figure 2.7.1.

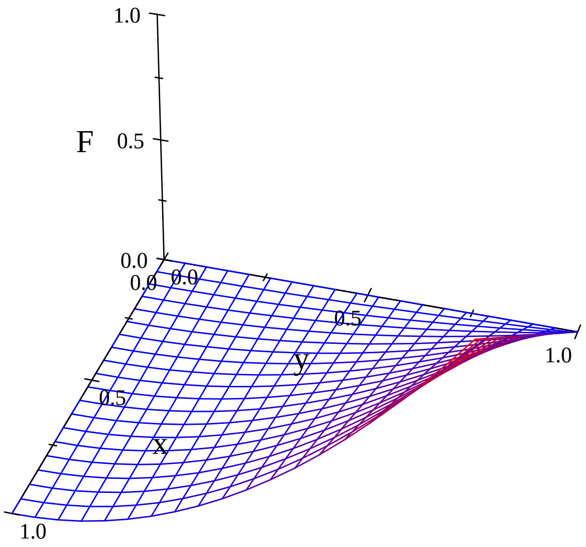

Figure 2.7.1: Graph of the joint distribution function $F_{X,Y}(x, y) = xy^2$ for $0 \le x \le 1$ and $0 \le y \le 1$ in Example 2.7.3.

▮

Theorem 2.7.3 thus tells us that the joint cdf $F_{X,Y}$ is very useful indeed. Not only does it tell us about the relationship of X to Y, but it also contains all the information about the marginal distributions of X and of Y.

We will see in the next subsections that joint probability functions, and joint density functions, similarly contain information about both the relationship of X and Y and the marginal distributions of X and Y.

2.7.3 Joint Probability Functions

Suppose X and Y are both *discrete* random variables. Then we can define a joint probability function for X and Y, as follows.

Definition 2.7.3 Let X and Y be discrete random variables. Then their joint probability function, $p_{X,Y}$, is a function from R^2 to R^1, defined by

$$p_{X,Y}(x, y) = P(X = x\,,\; Y = y).$$

Consider the following example.

EXAMPLE 2.7.4 (*Examples 2.7.1 and 2.7.2 continued*)
Again, let $X \sim$ Bernoulli$(1/2)$, $Y_1 = X$, and $Y_2 = 1 - X$. Then we see that

$$p_{X,Y_1}(x, y) = P(X = x\,,\ Y_1 = y) = \begin{cases} 1/2 & x = y = 1 \\ 1/2 & x = y = 0 \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand,

$$p_{X,Y_2}(x, y) = P(X = x\,,\ Y_2 = y) = \begin{cases} 1/2 & x = 1,\ y = 0 \\ 1/2 & x = 0,\ y = 1 \\ 0 & \text{otherwise.} \end{cases}$$

We thus see that p_{X,Y_1} and p_{X,Y_2} are two simple functions that are easy to work with and that clearly describe the relationships between X and Y_1 and between X and Y_2. Hence, for pairs of discrete random variables, joint probability functions are usually the best way to describe their relationships. ▮

Once we know the joint probability function $p_{X,Y}$, the marginal probability functions of X and Y are easily obtained.

Theorem 2.7.4 Let X and Y be two discrete random variables, with joint probability function $p_{X,Y}$. Then the probability function p_X of X can be computed as

$$p_X(x) = \sum_y p_{X,Y}(x, y).$$

Similarly, the probability function p_Y of Y can be computed as

$$p_Y(y) = \sum_x p_{X,Y}(x, y).$$

PROOF Using additivity of P, we have that

$$p_X(x) = P(X = x) = \sum_y P(X = x\,,\ Y = y) = \sum_y p_{X,Y}(x, y),$$

as claimed. Similarly,

$$p_Y(y) = P(Y = y) = \sum_x P(X = x,\ Y = y) = \sum_x p_{X,Y}(x, y). \ ▮$$

EXAMPLE 2.7.5
Suppose the joint probability function of X and Y is given by

$$p_{X,Y}(x, y) = \begin{cases} 1/7 & x = 5, y = 0 \\ 1/7 & x = 5, y = 3 \\ 1/7 & x = 5, y = 4 \\ 3/7 & x = 8, y = 0 \\ 1/7 & x = 8, y = 4 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} p_X(5) &= \sum_y p_{X,Y}(5, y) = p_{X,Y}(5, 0) + p_{X,Y}(5, 3) + p_{X,Y}(5, 4) \\ &= \frac{1}{7} + \frac{1}{7} + \frac{1}{7} = \frac{3}{7}, \end{aligned}$$

while

$$p_X(8) = \sum_y p_{X,Y}(8, y) = p_{X,Y}(8, 0) + p_{X,Y}(8, 4) = \frac{3}{7} + \frac{1}{7} = \frac{4}{7}.$$

Similarly,

$$p_Y(4) = \sum_x p_{X,Y}(x, 4) = p_{X,Y}(5, 4) + p_{X,Y}(8, 4) = \frac{1}{7} + \frac{1}{7} = \frac{2}{7},$$

etc.

Note that in such a simple context it is possible to tabulate the joint probability function in a table, as illustrated below for $p_{X,Y}$, p_X, and p_Y of this example.

	$Y = 0$	$Y = 3$	$Y = 4$	
$X = 5$	$1/7$	$1/7$	$1/7$	$3/7$
$X = 8$	$3/7$	0	$1/7$	$4/7$
	$4/7$	$1/7$	$2/7$	

Summing the rows and columns and placing the totals in the margins gives the marginal distributions of X and Y. ▮

2.7.4 Joint Density Functions

If X and Y are *continuous* random variables, then clearly $p_{X,Y}(x, y) = 0$ for all x and y. Hence, joint probability functions are not useful in this case. On the other hand, we shall see here that if X and Y are *jointly absolutely continuous*, then their relationship may be usefully described by a joint density function.

Definition 2.7.4 Let $f : R^2 \to R^1$ be a function. Then f is a *joint density function* if $f(x, y) \geq 0$ for all x and y, and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$.

Definition 2.7.5 Let X and Y be random variables. Then X and Y are *jointly absolutely continuous* if there is a joint density function f, such that

$$P(a \leq X \leq b\,,\ c \leq Y \leq d) = \int_c^d \int_a^b f(x, y)\, dx\, dy,$$

for all $a \leq b\,, c \leq d\,.$

Consider the following example.

EXAMPLE 2.7.6

Let X and Y be jointly absolutely continuous, with joint density function f given by

$$f(x, y) = \begin{cases} 4x^2y + 2y^5 & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

We first verify that f is indeed a density function. Clearly, $f(x, y) \ge 0$ for all x and y. Also,

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy &= \int_0^1\int_0^1 (4x^2y + 2y^5)\,dx\,dy = \int_0^1 \left(\frac{4}{3}y + 2y^5\right) dy \\ &= \frac{4}{3}\frac{1}{2} + 2\frac{1}{6} = \frac{2}{3} + \frac{1}{3} = 1. \end{aligned}$$

Hence, f is a joint density function. In Figure 2.7.2, we have plotted the function f, which gives a surface over the unit square.

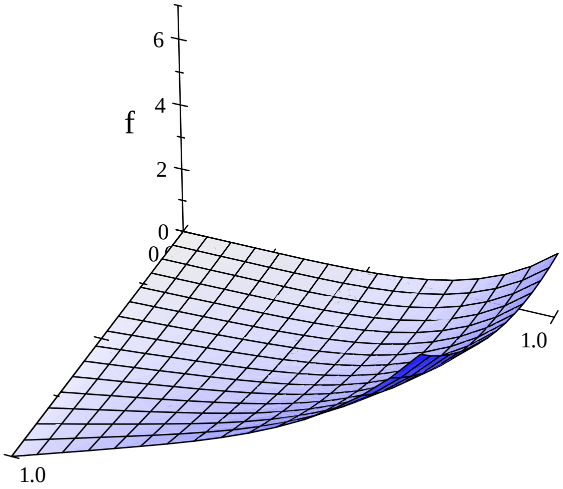

Figure 2.7.2: A plot of the density f in Example 2.7.6.

We next compute $P(0.5 \le X \le 0.7\,,\ 0.2 \le Y \le 0.9)$. Indeed, we have

$$\begin{aligned}
&P(0.5 \le X \le 0.7\,,\ 0.2 \le Y \le 0.9)\\
&= \int_{0.2}^{0.9}\int_{0.5}^{0.7}(4x^2y + 2y^5)\,dx\,dy\\
&= \int_{0.2}^{0.9}\left[\left(\frac{4}{3}((0.7)^3 - (0.5)^3\right)y + 2y^5(0.7-0.5)\right]dy\\
&= \frac{4}{3}\left((0.7)^3 - (0.5)^3\right)\frac{1}{2}((0.9)^2 - (0.2)^2) + \frac{2}{6}((0.9)^6 - (0.2)^6)(0.7-0.5)\\
&= \frac{2}{3}((0.7)^3 - (0.5)^3)((0.9)^2 - (0.2)^2) + \frac{1}{3}((0.9)^6 - (0.2)^6)(0.7-0.5) \doteq 0.147.
\end{aligned}$$

Other probabilities can be computed similarly. ∎

Once we know a joint density $f_{X,Y}$, then computing the marginal densities of X and Y is very easy, as the following theorem shows.

Theorem 2.7.5 Let X and Y be jointly absolutely continuous random variables, with joint density function $f_{X,Y}$. Then the (marginal) density f_X of X satisfies

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

for all $x \in R^1$. Similarly, the (marginal) density f_Y of Y satisfies

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

for all $y \in R^1$.

PROOF We need to show that, for $a \le b$, $P(a \le X \le b) = \int_a^b f_X(x)\,dx = \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dy\,dx$. Now, we always have $-\infty < Y < \infty$. Hence, using continuity of P, we have that $P(a \le X \le b) = P(a \le X \le b\,, -\infty < Y < \infty)$ and

$$\begin{aligned}
&P(a \le X \le b\,, -\infty < Y < \infty)\\
&= \lim_{\substack{c\to-\infty\\ d\to\infty}} P(a \le X \le b, c \le Y \le d) = \lim_{\substack{c\to-\infty\\ d\to\infty}} \int_c^d \int_a^b f(x, y)\,dx\,dy\\
&= \lim_{\substack{c\to-\infty\\ d\to\infty}} \int_a^b \int_c^d f(x, y)\,dy\,dx = \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dy\,dx,
\end{aligned}$$

as claimed. The result for f_Y follows similarly. ∎

EXAMPLE 2.7.7 (*Example 2.7.6 continued*)
Let X and Y again have joint density

$$f_{X,Y}(x, y) = \begin{cases} 4x^2y + 2y^5 & 0 \le x \le 1\,,\ 0 \le y \le 1\\ 0 & \text{otherwise.}\end{cases}$$

Then by Theorem 2.7.5, for $0 \le x \le 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_0^1 (4x^2y + 2y^5)\,dy \ = \ 2x^2 + (1/3),$$

while for $x < 0$ or $x > 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{\infty} 0\,dy = 0.$$

Similarly, for $0 \le y \le 1$,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \int_0^1 (4x^2y + 2y^5)\,dx = \frac{4}{3}y + 2y^5,$$

while for $y < 0$ or $y > 1$, $f_Y(y) = 0$. ▮

EXAMPLE 2.7.8

Suppose X and Y are jointly absolutely continuous, with joint density

$$f_{X,Y}(x,y) = \begin{cases} 120x^3y & x \ge 0,\ y \ge 0,\ x+y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then the region where $f_{X,Y}(x,y) > 0$ is a *triangle*, as depicted in Figure 2.7.3.

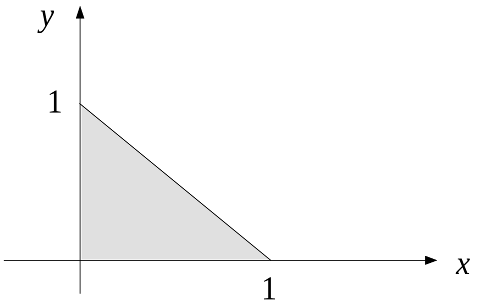

Figure 2.7.3: Region of the plane where the density $f_{X,Y}$ in Example 2.7.8 is positive.

We check that

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy &= \int_0^1\int_0^{1-x} 120x^3y\,dy\,dx = \int_0^1 120x^3\frac{(1-x)^2}{2}\,dx \\ &= \int_0^1 60(x^3 - 2x^4 + x^5)\,dx = 60(\frac{1}{4} - 2\frac{1}{5} + \frac{1}{6}) \\ &= 15 - 2(12) + 10 = 1,\end{aligned}$$

so that $f_{X,Y}$ is indeed a joint density function. We then compute that, for example,

$$f_X(x) = \int_0^{1-x} 120x^3y\,dy = 120x^3\frac{(1-x)^2}{2} = 60(x^3 - 2x^4 + x^5)$$

for $0 \le x \le 1$ (with $f_X(x) = 0$ for $x < 0$ or $x > 1$). ∎

EXAMPLE 2.7.9 *Bivariate Normal*$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ *Distribution*
Let $\mu_1, \mu_2, \sigma_1, \sigma_2$, and ρ be real numbers, with $\sigma_1, \sigma_2 > 0$ and $-1 \le \rho \le 1$. Let X and Y have joint density given by

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\begin{array}{c} \left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - \\ 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) \end{array} \right] \right\}$$

for $x \in R^1$, $y \in R^1$. We say that X and Y have the *Bivariate Normal*$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ *distribution*.

It can be shown (see Problem 2.7.13) that $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. Hence, X and Y are each normally distributed. The parameter ρ measures the degree of the *relationship* that exists between X and Y (see Problem 3.3.17) and is called the *correlation*. In particular, X and Y are independent (see Section 2.8.3), and so unrelated, if and only if $\rho = 0$ (see Problem 2.8.21).

Figure 2.7.4 is a plot of the *standard bivariate normal* density, given by setting $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 1, \sigma_2 = 1$, and $\rho = 0$. This is a bell-shaped surface in R^3 with its peak at the point $(0, 0)$ in the xy-plane. The graph of the general Bivariate Normal$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ distribution is also a bell-shaped surface, but the peak is at the point (μ_1, μ_2) in the xy-plane and the shape of the bell is controlled by σ_1, σ_2, and ρ.

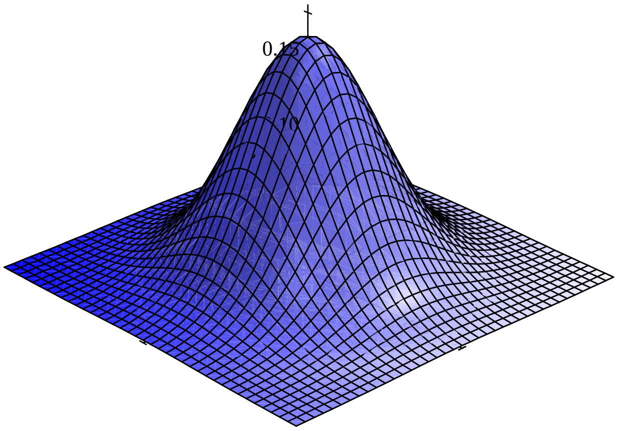

Figure 2.7.4: A plot of the standard bivariate normal density function.

It can be shown (see Problem 2.9.16) that, when Z_1, Z_2 are independent random variables, both distributed $N(0, 1)$, and we put

$$X = \mu_1 + \sigma_1 Z_1, \quad Y = \mu_2 + \sigma_2\left(\rho Z_1 + (1-\rho^2)^{1/2} Z_2\right), \tag{2.7.1}$$

then $(X, Y) \sim$ Bivariate Normal$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. This relationship can be quite useful in establishing various properties of this distribution. We can also write an analogous version $Y = \mu_2 + \sigma_2 Z_1$, $X = \mu_1 + \sigma_1(\rho Z_1 + \left(1 - \rho^2\right)^{1/2} Z_2)$ and obtain the same distributional result.

The bivariate normal distribution is one of the most commonly used bivariate distributions in applications. For example, if we randomly select an individual from a population and measure his weight X and height Y, then a bivariate normal distribution will often provide a reasonable description of the joint distribution of these variables. ∎

Joint densities can also be used to compute probabilities of more general regions, as the following result shows. (We omit the proof. The special case $B = [a, b] \times [c, d]$ corresponds directly to the definition of $f_{X,Y}$.)

Theorem 2.7.6 Let X and Y be jointly absolutely continuous random variables, with joint density $f_{X,Y}$, and let $B \subseteq R^2$ be any region. Then

$$P\big((X, Y) \in B\big) = \int\int_B f(x, y)\, dx\, dy.$$

The previous discussion has centered around having just *two* random variables, X and Y. More generally, we may consider n random variables $X_1, \ldots, X_n$. If the random variables are all discrete, then we can further define a joint probability function $p_{X_1,\ldots,X_n} : R^n \to [0, 1]$ by $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$. If the random variables are jointly absolutely continuous, then we can define a joint density function $f_{X_1,\ldots,X_n} : R^n \to [0, 1]$ so that

$$\begin{aligned} &P(a_1 \le X_1 \le b_1, \ldots, a_n \le X_n \le b_n) \\ &= \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\, dx_1 \cdots dx_n, \end{aligned}$$

whenever $a_i \le b_i$ for all i.

Summary of Section 2.7

- It is often important to keep track of the joint probabilities of two random variables, X and Y.
- Their joint cumulative distribution function is given by $F_{X,Y}(x, y) = P(X \le x,\ Y \le y)$.
- If X and Y are discrete, then their joint probability function is given by $p_{X,Y}(x, y) = P(X = x,\ Y = y)$.
- If X and Y are absolutely continuous, then their joint density function $f_{X,Y}(x, y)$ is such that $P(a \le X \le b,\ c \le Y \le d) = \int_c^d \int_a^b f_{X,Y}(x, y)\, dx\, dy$.
- The marginal density of X and Y can be computed from any of $F_{X,Y}$, or $p_{X,Y}$, or $f_{X,Y}$.
- An important example of a joint distribution is the bivariate normal distribution.

EXERCISES

2.7.1 Let $X \sim$ Bernoulli$(1/3)$, and let $Y = 4X - 2$. Compute the joint cdf $F_{X,Y}$.

2.7.2 Let $X \sim$ Bernoulli$(1/4)$, and let $Y = -7X$. Compute the joint cdf $F_{X,Y}$.

2.7.3 Suppose

$$p_{X,Y}(x, y) = \begin{cases} 1/5 & x = 2,\ y = 3 \\ 1/5 & x = 3,\ y = 2 \\ 1/5 & x = -3,\ y = -2 \\ 1/5 & x = -2,\ y = -3 \\ 1/5 & x = 17,\ y = 19 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute p_X.
(b) Compute p_Y.
(c) Compute $P(Y > X)$.
(d) Compute $P(Y = X)$.
(e) Compute $P(XY < 0)$.

2.7.4 For each of the following joint density functions $f_{X,Y}$, find the value of C and compute $f_X(x)$, $f_Y(y)$, and $P(X \le 0.8\,, Y \le 0.6)$.
(a)

$$f_{X,Y}(x, y) = \begin{cases} 2x^2y + Cy^5 & 0 \le x \le 1\,,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(b)

$$f_{X,Y}(x, y) = \begin{cases} C(xy + x^5y^5) & 0 \le x \le 1\,,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(c)

$$f_{X,Y}(x, y) = \begin{cases} C(xy + x^5y^5) & 0 \le x \le 4\,,\ 0 \le y \le 10 \\ 0 & \text{otherwise.} \end{cases}$$

(d)

$$f_{X,Y}(x, y) = \begin{cases} Cx^5y^5 & 0 \le x \le 4\,,\ 0 \le y \le 10 \\ 0 & \text{otherwise.} \end{cases}$$

2.7.5 Prove that $F_{X,Y}(x, y) \le \min(F_X(x), F_Y(y))$.

2.7.6 Suppose $P(X = x,\ Y = y) = 1/8$ for $x = 3, 5$ and $y = 1, 2, 4, 7$, otherwise $P(X = x,\ Y = y) = 0$. Compute each of the following.
(a) $F_{X,Y}(x, y)$ for all $x, y \in R^1$
(b) $p_{X,Y}(x, y)$ for all $x, y \in R^1$
(c) $p_X(x)$ for all $x \in R^1$
(d) $p_Y(y)$ for all $x \in R^1$
(e) The marginal cdf $F_X(x)$ for all $x \in R^1$
(f) The marginal cdf $F_Y(y)$ for all $y \in R^1$

2.7.7 Let X and Y have joint density $f_{X,Y}(x, y) = c \sin(xy)$ for $0 < x < 1$ and $0 < y < 2$, otherwise $f_{X,Y}(x, y) = 0$, for appropriate constant $c > 0$ (which cannot be computed explicitly). In terms of c, compute each of the following.
(a) The marginal density $f_X(x)$ for all $x \in R^1$
(b) The marginal density $f_Y(y)$ for all $y \in R^1$

2.7.8 Let X and Y have joint density $f_{X,Y}(x, y) = (x^2 + y)/36$ for $-2 < x < 1$ and $0 < y < 4$, otherwise $f_{X,Y}(x, y) = 0$. Compute each of the following.
(a) The marginal density $f_X(x)$ for all $x \in R^1$
(b) The marginal density $f_Y(y)$ for all $y \in R^1$
(c) $P(Y < 1)$
(d) The joint cdf $F_{X,Y}(x, y)$ for all $x, y \in R^1$

2.7.9 Let X and Y have joint density $f_{X,Y}(x, y) = (x^2 + y)/4$ for $0 < x < y < 2$, otherwise $f_{X,Y}(x, y) = 0$. Compute each of the following.
(a) The marginal density $f_X(x)$ for all $x \in R^1$
(b) The marginal density $f_Y(y)$ for all $y \in R^1$
(c) $P(Y < 1)$

2.7.10 Let X and Y have the Bivariate-Normal(3, 5, 2, 4, 1/2) distribution.
(a) Specify the marginal distribution of X.
(b) Specify the marginal distribution of Y.
(c) Are X and Y independent? Why or why not?

PROBLEMS

2.7.11 Let $X \sim$ Exponential(λ), and let $Y = X^3$. Compute the joint cdf, $F_{X,Y}(x, y)$.

2.7.12 Let $F_{X,Y}$ be a joint cdf. Prove that for all $y \in R^1$, $\lim_{x\to-\infty} F_{X,Y}(x, y) = 0$.

2.7.13 Let X and Y have the Bivariate Normal($\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$) distribution, as in Example 2.7.9. Prove that $X \sim N(\mu_1, \sigma_1^2)$, by proving that

$$\int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}.$$

2.7.14 Suppose that the joint density $f_{X,Y}$ is given by $f_{X,Y}(x, y) = Cye^{-xy}$ for $0 < x < 1$, $0 < y < 1$ and is 0 otherwise.

(a) Determine C so that $f_{X,Y}$ is a density.

(b) Compute $P\left(1/2 < X < 1, 1/2 < Y < 1\right)$.

(c) Compute the marginal densities of X and Y.

2.7.15 Suppose that the joint density $f_{X,Y}$ is given by $f_{X,Y}(x, y) = Cye^{-xy}$ for $0 < x < y < 1$ and is 0 otherwise.

(a) Determine C so that $f_{X,Y}$ is a density.

(b) Compute $P\left(1/2 < X < 1, 1/2 < Y < 1\right)$.

(c) Compute the marginal densities of X and Y.

2.7.16 Suppose that the joint density $f_{X,Y}$ is given by $f_{X,Y}(x, y) = Ce^{-(x+y)}$ for $0 < x < y < \infty$ and is 0 otherwise.

(a) Determine C so that $f_{X,Y}$ is a density.

(b) Compute the marginal densities of X and Y.

2.7.17 (*Dirichlet*$(\alpha_1, \alpha_2, \alpha_3)$ *distribution*) Let (X_1, X_2) have the joint density

$$f_{X_1,X_2}(x_1, x_2) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)\,\Gamma(\alpha_3)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} (1 - x_1 - x_2)^{\alpha_3 - 1}$$

for $x_1 \geq 0, x_2 \geq 0$, and $0 \leq x_1 + x_2 \leq 1$. A Dirichlet distribution is often applicable when X_1, X_2, and $1 - X_1 - X_2$ correspond to random proportions.

(a) Prove that f_{X_1,X_2} is a density. (Hint: Sketch the region where f_{X_1,X_2} is nonnegative, integrate out x_1 first by making the transformation $u = x_1/(1 - x_2)$ in this integral, and use (2.4.10) from Problem 2.4.24.)

(b) Prove that $X_1 \sim \text{Beta}(\alpha_1, \alpha_2 + \alpha_3)$ and $X_2 \sim \text{Beta}(\alpha_2, \alpha_1 + \alpha_3)$.

2.7.18 (*Dirichlet*$(\alpha_1, \ldots, \alpha_{k+1})$ *distribution*) Let $(X_1, \ldots, X_k)$ have the joint density

$$\begin{aligned} &f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \\ &= \frac{\Gamma(\alpha_1 + \cdots + \alpha_{k+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{k+1})} x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} (1 - x_1 - \cdots - x_k)^{\alpha_{k+1} - 1} \end{aligned}$$

for $x_i \geq 0, i = 1, \ldots, k$, and $0 \leq x_1 + \cdots + x_k \leq 1$. Prove that $f_{X_1,\ldots,X_k}$ is a density. (Hint: Problem 2.7.17.)

CHALLENGES

2.7.19 Find an example of two random variables X and Y and a function $h : R^1 \to R^1$, such that $F_X(x) > 0$ and $F_Y(x) > 0$ for all $x \in R^1$, but $\lim_{x \to \infty} F_{X,Y}(x, h(x)) = 0$.

DISCUSSION TOPICS

2.7.20 What are examples of pairs of real-life random quantities that have interesting relationships? (List as many as you can, and describe each relationship as well as you can.)

2.8 Conditioning and Independence

Let X and Y be two random variables. Suppose we know that $X = 5$. What does that tell us about Y? Depending on the relationship between X and Y, that may tell us everything about Y (e.g., if $Y = X$), or nothing about Y. Usually, the answer will be between these two extremes, and the knowledge that $X = 5$ will change the probabilities for Y somewhat.

2.8.1 Conditioning on Discrete Random Variables

Suppose X is a discrete random variable, with $P(X = 5) > 0$. Let $a < b$, and suppose we are interested in the conditional probability $P(a < Y \leq b \,|\, X = 5)$. Well, we already know how to compute such conditional probabilities. Indeed, by (1.5.1),

$$P(a < Y \leq b \mid X = 5) = \frac{P(a < Y \leq b\,,\ X = 5)}{P(X = 5)},$$

provided that $P(X = 5) > 0$. This prompts the following definition.

Definition 2.8.1 Let X and Y be random variables, and suppose that $P(X = x) > 0$. The *conditional distribution* of Y, given that $X = x$, is the probability distribution assigning probability

$$\frac{P(Y \in B\,, X = x)}{P(X = x)}$$

to each event $Y \in B$. In particular, it assigns probability

$$\frac{P(a < Y \leq b\,, X = x)}{P(X = x)}$$

to the event that $a < Y \leq b$.

EXAMPLE 2.8.1

Suppose as in Example 2.7.5 that X and Y have joint probability function

$$p_{X,Y}(x, y) = \begin{cases} 1/7 & x = 5,\ y = 0 \\ 1/7 & x = 5,\ y = 3 \\ 1/7 & x = 5,\ y = 4 \\ 3/7 & x = 8,\ y = 0 \\ 1/7 & x = 8,\ y = 4 \\ 0 & \text{otherwise.} \end{cases}$$

We compute $P(Y = 4 \mid X = 8)$ as

$$P(Y = 4 \mid X = 8) = \frac{P(Y = 4\,,\ X = 8)}{P(X = 8)} = \frac{1/7}{(3/7) + (1/7)} = \frac{1/7}{4/7} = 1/4.$$

On the other hand,

$$P(Y = 4 \mid X = 5) = \frac{P(Y = 4\,,\ X = 5)}{P(X = 5)} = \frac{1/7}{(1/7) + (1/7) + (1/7)} = \frac{1/7}{3/7} = 1/3.$$

Thus, depending on the value of X, we obtain different probabilities for Y. ▮

Generalizing from the above example, we see that if X and Y are discrete, then

$$P(Y = y \mid X = x) = \frac{P(Y = y\,,\ X = x)}{P(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_{X,Y}(x, y)}{\sum_z p_{X,Y}(x, z)}.$$

This prompts the following definition.

Definition 2.8.2 Suppose X and Y are two discrete random variables. Then the *conditional probability function* of Y, given X, is the function $p_{Y|X}$ defined by

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{\sum_z p_{X,Y}(x, z)} = \frac{p_{X,Y}(x, y)}{p_X(x)},$$

defined for all $y \in R^1$ and all x with $p_X(x) > 0$.

2.8.2 Conditioning on Continuous Random Variables

If X is continuous, then we will have $P(X = x) = 0$. In this case, Definitions 2.8.1 and 2.8.2 cannot be used because we cannot divide by 0. So how can we condition on $X = x$ in this case?

One approach is suggested by instead conditioning on $x - \epsilon < X \le x + \epsilon$, where $\epsilon > 0$ is a very small number. Even if X is continuous, we might still have $P(x - \epsilon \le X \le x + \epsilon) > 0$. On the other hand, if ϵ is very small and $x - \epsilon \le X \le x + \epsilon$, then X must be very close to x.

Indeed, suppose that X and Y are jointly absolutely continuous, with joint density function $f_{X,Y}$. Then

$$\begin{aligned} P(a \le Y \le b \mid x - \epsilon \le X \le x + \epsilon) &= \frac{P(a \le Y \le b, x - \epsilon \le X \le x + \epsilon)}{P(x - \epsilon \le X \le x + \epsilon)} \\ &= \frac{\int_a^b \int_{x-\epsilon}^{x+\epsilon} f_{X,Y}(t, y)\, dt\, dy}{\int_{-\infty}^{\infty} \int_{x-\epsilon}^{x+\epsilon} f_{X,Y}(t, y)\, dt\, dy}. \end{aligned}$$

In Figure 2.8.1, we have plotted the region $\{(x, y) : a \le y \le b\,,\ x - \epsilon < x \le x + \epsilon\}$ for (X, Y).

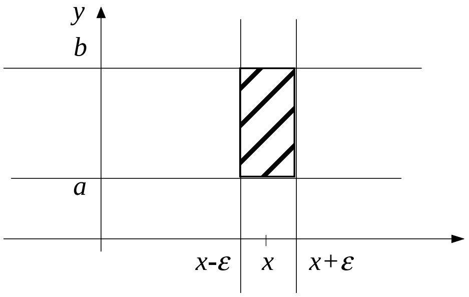

Figure 2.8.1: The shaded region is the set $\{(x, y) : a \le y \le b\,,\ x - \epsilon \le x \le x + \epsilon\}$.

Now, if ϵ is very small, then in the above integrals we will always have t very close to x. If $f_{X,Y}$ is a continuous function, then this implies that $f_{X,Y}(t, y)$ will be very close to $f_{X,Y}(x, y)$. We conclude that, if ϵ is very small, then

$$
\begin{aligned}
P(a \le Y \le b \,|\, x - \epsilon \le X \le x + \epsilon) &\approx \frac{\int_a^b \int_{x-\epsilon}^{x+\epsilon} f_{X,Y}(x, y)\, dt\, dy}{\int_{-\infty}^{\infty} \int_{x-\epsilon}^{x+\epsilon} f_{X,Y}(x, y)\, dt\, dy} \\
&= \frac{\int_a^b 2\epsilon\ f_{X,Y}(x, y)\, dy}{\int_{-\infty}^{\infty} 2\epsilon\ f_{X,Y}(x, y)\, dy} = \int_a^b \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, z)\, dz}\, dy.
\end{aligned}
$$

This suggests that the quantity

$$\frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, z)\, dz} = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

plays the role of a density, for the conditional distribution of Y, given that $X = x$. This prompts the following definitions.

Definition 2.8.3 Let X and Y be jointly absolutely continuous, with joint density function $f_{X,Y}$. The *conditional density* of Y, given $X = x$, is the function $f_{Y|X}(y \,|\, x)$, defined by

$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x, y)}{f_X(x)},$$

valid for all $y \in R^1$, and for all x such that $f_X(x) > 0$.

Definition 2.8.4 Let X and Y be jointly absolutely continuous, with joint density function $f_{X,Y}$. The *conditional distribution* of Y, given $X = x$, is defined by saying that

$$P(a \le Y \le b \,|\, X = x) = \int_a^b f_{Y|X}(y \,|\, x)\, dy,$$

when $a \le b$, with $f_{Y|X}$ as in Definition 2.8.3, valid for all x such that $f_X(x) > 0$.

EXAMPLE 2.8.2

Let X and Y have joint density

$$f_{X,Y}(x, y) = \begin{cases} 4x^2y + 2y^5 & 0 \le x \le 1\,,\ 0 \le y \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

as in Examples 2.7.6 and 2.7.7.

We know from Example 2.7.7 that

$$f_X(x) = \begin{cases} 2x^2 + (1/3) & 0 \le x \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

while

$$f_Y(y) = \begin{cases} \frac{4}{3}y + 2y^5 & 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let us now compute $P(0.2 \le Y \le 0.3 \,|\, X = 0.8)$. Using Definitions 2.8.4 and 2.8.3, we have

$$
\begin{aligned}
&P(0.2 \le Y \le 0.3 \,|\, X = 0.8)\\
&= \int_{0.2}^{0.3} f_{Y|X}(y \,|\, 0.8)\, dy = \frac{\int_{0.2}^{0.3} f_{X,Y}(0.8,\ y)\, dy}{f_X(0.8)} = \frac{\int_{0.2}^{0.3} \left(4\,(0.8)^2\, y + 2y^5\right)\, dy}{2\,(0.8)^2 + \frac{1}{3}}\\
&= \frac{\frac{4}{2}\,(0.8)^2\,((0.3)^2 - (0.2)^2) + \frac{2}{6}((0.3)^6 - (0.2)^6)}{2\ (0.8)^2 + \frac{1}{3}} = 0.0398.
\end{aligned}
$$

By contrast, if we compute the *unconditioned* (i.e., usual) probability that $0.2 \le Y \le 0.3$, we see that

$$
\begin{aligned}
P(0.2 \le Y \le 0.3) &= \int_{0.2}^{0.3} f_Y(y)\, dy = \int_{0.2}^{0.3} (\frac{4}{3}y + 2y^5)\, dy\\
&= \frac{4}{3}\frac{1}{2}((0.3)^2 - (0.2)^2) + \frac{2}{6}((0.3)^6 - (0.2)^6) = 0.0336.
\end{aligned}
$$

We thus see that conditioning on $X = 0.8$ increases the probability that $0.2 \le Y \le 0.3$, from about 0.0336 to about 0.0398. ∎

By analogy with Theorem 1.3.1, we have the following.

Theorem 2.8.1 (*Law of total probability, absolutely continuous random variable version*) Let X and Y be jointly absolutely continuous random variables, and let $a \le b$ and $c \le d$. Then

$$P(a \le X \le b,\ c \le Y \le d) = \int_c^d \int_a^b f_X(x)\ f_{Y|X}(y \,|\, x)\, dx\, dy.$$

More generally, if $B \subseteq R^2$ is any region, then

$$P\big((X, Y) \in B\big) = \int\int_B f_X(x)\ f_{Y|X}(y \,|\, x)\, dx\, dy.$$

PROOF By Definition 2.8.3,

$$f_X(x)\ f_{Y|X}(y \,|\, x) = f_{X,Y}(x, y).$$

Hence, the result follows immediately from Definition 2.7.4 and Theorem 2.7.6. ∎

2.8.3 Independence of Random Variables

Recall from Definition 1.5.2 that two events A and B are *independent* if $P(A \cap B) = P(A)\,P(B)$. We wish to have a corresponding definition of independence for random variables X and Y. Intuitively, independence of X and Y means that X and Y have no influence on each other, i.e., that the values of X make no change to the probabilities for Y (and vice versa).

The idea of the formal definition is that X and Y give rise to events, of the form "$a < X \leq b$" or "$Y \in B$," and we want all such events involving X to be independent of all such events involving Y. Specifically, our definition is the following.

Definition 2.8.5 Let X and Y be two random variables. Then X and Y are *independent* if, for all subsets B_1 and B_2 of the real numbers,

$$P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2).$$

That is, the events "$X \in B_1$" and "$Y \in B_2$" are independent events.

Intuitively, X and Y are independent if they have no influence on each other, as we shall see.

Now, Definition 2.8.5 is very difficult to work with. Fortunately, there is a much simpler characterization of independence.

Theorem 2.8.2 Let X and Y be two random variables. Then X and Y are independent if and only if

$$P(a \leq X \leq b, c \leq Y \leq d) = P(a \leq X \leq b)P(c \leq Y \leq d) \qquad (2.8.1)$$

whenever $a \leq b$ and $c \leq d$.

That is, X and Y are independent if and only if the events "$a \leq X \leq b$" and "$c \leq Y \leq d$" are independent events whenever $a \leq b$ and $c \leq d$.

We shall not prove Theorem 2.8.2 here, although it is similar in spirit to the proof of Theorem 2.5.1. However, we shall sometimes use (2.8.1) to check for the independence of X and Y.

Still, even (2.8.1) is not so easy to check directly. For discrete and for absolutely continuous distributions, easier conditions are available, as follows.

Theorem 2.8.3 Let X and Y be two random variables.
(a) If X and Y are discrete, then X and Y are independent if and only if their joint probability function $p_{X,Y}$ satisfies

$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$$

for all $x, y \in R^1$.
(b) If X and Y are jointly absolutely continuous, then X and Y are independent if and only if their joint density function $f_{X,Y}$ can be chosen to satisfy

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

for all $x, y \in R^1$.

PROOF (a) If X and Y are independent, then setting $a = b = x$ and $c = d = y$ in (2.8.1), we see that $P(X = x, Y = y) = P(X = x)P(Y = y)$. Hence, $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$.

Conversely, if $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$ for all x and y, then

$$\begin{aligned}
&P(a \le X \le b, c \le Y \le d)\\
&= \sum_{a\le x\le b}\ \sum_{c\le y\le d} p_{X,Y}(x, y) = \sum_{a\le x\le b}\ \sum_{c\le y\le d} p_X(x)\, p_Y(y)\\
&= \left(\sum_{a\le x\le b} p_X(x)\right)\left(\sum_{c\le y\le d} p_Y(y)\right) = P(a \le X \le b)\ P(c \le Y \le d).
\end{aligned}$$

This completes the proof of (a).

(b) If $f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$ for all x and y, then

$$\begin{aligned}
&P(a \le X \le b, c \le Y \le d)\\
&= \int_a^b \int_c^d f_{X,Y}(x, y)\, dy\, dx = \int_a^b \int_c^d f_X(x)\ f_Y(y)\, dy\, dx\\
&= \left(\int_a^b f_X(x)\, dx\right)\left(\int_c^d f_Y(y)\, dy\right) = P(a \le X \le b)\ P(c \le Y \le d).
\end{aligned}$$

This completes the proof of the "if" part of (b). The proof of the "only if" part of (b) is more technical, and we do not include it here. ▮

EXAMPLE 2.8.3

Let X and Y have, as in Example 2.7.6, joint density

$$f_{X,Y}(x, y) = \begin{cases} 4x^2y + 2y^5 & 0 \le x \le 1\,,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and so, as derived in as in Example 2.7.7, marginal densities

$$f_X(x) = \begin{cases} 2x^2 + (1/3) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_Y(y) = \begin{cases} \frac{4}{3}y + 2y^5 & 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then we compute that

$$f_X(x)\ f_Y(y) = \begin{cases} (2x^2 + (1/3))\,(\frac{4}{3}y + 2y^5) & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

We therefore see that $f_X(x)\ f_Y(y) \ne f_{X,Y}(x, y)$. Hence, X and Y are *not* independent. ▮

EXAMPLE 2.8.4

Let X and Y have joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{8080}\,(12xy^2 + 6x + 4y^2 + 2) & 0 \le x \le 6,\ 3 \le y \le 5 \\ 0 & \text{otherwise.} \end{cases}$$

We compute the marginal densities as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \begin{cases} \frac{1}{60} + \frac{1}{20}\,x & 0 \le x \le 6 \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx = \begin{cases} \frac{3}{202} + \frac{3}{101}\,y^2 & 3 \le y \le 5 \\ 0 & \text{otherwise.} \end{cases}$$

Then we compute that

$$f_X(x)\,f_Y(y) = \begin{cases} (\frac{1}{60} + \frac{1}{20}\,x)\,(\frac{3}{202} + \frac{3}{101}\,y^2) & 0 \le x \le 6\,,\ 3 \le y \le 5 \\ 0 & \text{otherwise.} \end{cases}$$

Multiplying this out, we see that $f_X(x)\,f_Y(y) = f_{X,Y}(x,y)$. Hence, X and Y *are* independent in this case. ▮

Combining Theorem 2.8.3 with Definitions 2.8.2 and 2.8.3, we immediately obtain the following result about independence. It says that independence of random variables is the same as saying that conditioning on one has no effect on the other, which corresponds to an intuitive notion of independence.

Theorem 2.8.4 Let X and Y be two random variables.
(a) If X and Y are discrete, then X and Y are independent if and only if $p_{Y|X}(y\,|\,x) = p_Y(y)$, for every $x, y \in R^1$.
(b) If X and Y are jointly absolutely continuous, then X and Y are independent if and only if $f_{Y|X}(y\,|\,x) = f_Y(y)$, for every $x, y \in R^1$.

While Definition 2.8.5 is quite difficult to work with, it does provide the easiest way to prove one very important property of independence, as follows.

Theorem 2.8.5 Let X and Y be independent random variables. Let $f, g : R^1 \to R^1$ be any two functions. Then the random variables $f(X)$ and $g(Y)$ are also independent.

PROOF Using Definition 2.8.5, we compute that

$$\begin{aligned} P\,(f(X) \in B_1,\ g(Y) \in B_2) &= P\left(X \in f^{-1}(B_1),\ Y \in g^{-1}(B_2)\right) \\ &= P\left(X \in f^{-1}(B_1)\right)\ P\left(Y \in g^{-1}(B_2)\right) \\ &= P\,(f(X) \in B_1)\ P\,(g(Y) \in B_2)\,. \end{aligned}$$

(Here $f^{-1}(B_1) = \{x \in R^1 : f(x) \in B_1\}$ and $g^{-1}(B_2) = \{y \in R^1 : g(y) \in B_2\}$.) Because this is true for any B_1 and B_2, we see that $f(X)$ and $g(Y)$ are independent. ∎

Suppose now that we have n random variables $X_1, \ldots, X_n$. The random variables are *independent* if and only if the collection of events $\{a_i \le X_i \le b_i\}$ are independent, whenever $a_i \le b_i$, for all $i = 1, 2, \ldots, n$. Generalizing Theorem 2.8.3, we have the following result.

Theorem 2.8.6 Let $X_1, \ldots, X_n$ be a collection of random variables.
(a) If $X_1, \ldots, X_n$ are discrete, then $X_1, \ldots, X_n$ are independent if and only if their joint probability function $p_{X_1,\ldots,X_n}$ satisfies

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$$

for all $x_1, \ldots, x_n \in R^1$.
(b) If $X_1, \ldots, X_n$ are jointly absolutely continuous, then $X_1, \ldots, X_n$ are independent if and only if their joint density function $f_{X_1,\ldots,X_n}$ can be chosen to satisfy

$$f_{X_1,\ldots,X_n}(x, y) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for all $x_1, \ldots, x_n \in R^1$.

A particularly common case in statistics is the following.

Definition 2.8.6 A collection $X_1, \ldots, X_n$ of random variables is *independent and identically distributed* (or *i.i.d.*) if the collection is independent and if, furthermore, each of the n variables has the same distribution. The i.i.d. sequence $X_1, \ldots, X_n$ is also referred to as a *sample* from the common distribution.

In particular, if a collection $X_1, \ldots, X_n$ of random variables is i.i.d. and *discrete*, then each of the probability functions p_{X_i} is the same, so that $p_{X_1}(x) = p_{X_2}(x) = \cdots = p_{X_n}(x) \equiv p(x)$, for all $x \in R^1$. Furthermore, from Theorem 2.8.6(a), it follows that

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \cdots p_{X_n}(x_n) = p(x_1)p(x_2) \cdots p(x_n)$$

for all $x_1, \ldots, x_n \in R^1$.

Similarly, if a collection $X_1, \ldots, X_n$ of random variables is i.i.d. and *jointly absolutely continuous*, then each of the density functions f_{X_i} is the same, so that $f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_n}(x) \equiv f(x)$, for all $x \in R^1$. Furthermore, from Theorem 2.8.6(b), it follows that

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_n}(x_n) = f(x_1)\, f(x_2) \cdots f(x_n)$$

for all $x_1, \ldots, x_n \in R^1$.

We now consider an important family of discrete distributions that arise via sampling.

EXAMPLE 2.8.5 *Multinomial Distributions*
Suppose we have a response s that can take three possible values — for convenience, labelled 1, 2, and 3 — with the probability distribution

$$P(s=1)=\theta_1,\ P(s=2)=\theta_2,\ P(s=3)=\theta_3$$

so that each $\theta_i \geq 0$ and $\theta_1+\theta_2+\theta_3=1$. As a simple example, consider a bowl of chips of which a proportion θ_i of the chips are labelled i (for $i = 1, 2, 3$). If we randomly draw a chip from the bowl and observe its label s, then $P(s=i)=\theta_i$. Alternatively, consider a population of students at a university of which a proportion θ_1 live on campus (denoted by $s = 1$), a proportion θ_2 live off-campus with their parents (denoted by $s = 2$), and a proportion θ_3 live off-campus independently (denoted by $s = 3$). If we randomly draw a student from this population and determine s for that student, then $P(s=i)=\theta_i$.

We can also write
$$P(s=i)=\theta_1^{I_{\{1\}}(i)}\theta_2^{I_{\{2\}}(i)}\theta_3^{I_{\{3\}}(i)}$$

for $i \in \{1, 2, 3\}$, where $I_{\{j\}}$ is the indicator function for $\{j\}$. Therefore, if $(s_1,\ldots,s_n)$ is a sample from the distribution on $\{1, 2, 3\}$ given by the θ_i, Theorem 2.8.6(a) implies that the joint probability function for the sample equals

$$P(s_1=k_1,\ldots,s_n=k_n)=\prod_{j=1}^{n}\theta_1^{I_{\{1\}}(k_j)}\theta_2^{I_{\{2\}}(k_j)}\theta_3^{I_{\{3\}}(k_j)}=\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3} \tag{2.8.2}$$

where $x_i=\sum_{j=1}^{n} I_{\{i\}}\left(k_j\right)$ is equal to the number of i's in $(k_1,\ldots,k_n)$.

Now, based on the sample $(s_1,\ldots,s_n)$, define the random variables

$$X_i=\sum_{j=1}^{n} I_{\{i\}}\left(s_j\right)$$

for $i = 1, 2$, and 3. Clearly, X_i is the number of i's observed in the sample and we always have $X_i \in \{0, 1, \ldots, n\}$ and $X_1+X_2+X_3=n$. We refer to the X_i as the *counts* formed from the sample.

For (x_1,x_2,x_3) satisfying $x_i \in \{0, 1, \ldots, n\}$ and $x_1+x_2+x_3=n$, (2.8.2) implies that the joint probability function for (X_1, X_2, X_3) is given by

$$\begin{aligned} p_{(X_1,X_2,X_3)}(x_1,x_2,x_3) &= P(X_1=x_1, X_2=x_2, X_3=x_3)\\ &= C(x_1,x_2,x_3)\,\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3} \end{aligned}$$

where $C(x_1,x_2,x_3)$ equals the number of samples $(s_1,\ldots,s_n)$ with x_1 of its elements equal to 1, x_2 of its elements equal to 2, and x_3 of its elements equal to 3. To calculate $C(x_1,x_2,x_3)$, we note that there are $\binom{n}{x_1}$ choices for the places of the 1's in the sample sequence, $\binom{n-x_1}{x_2}$ choices for the places of the 2's in the sequence, and finally $\binom{n-x_1-x_2}{x_3}=1$ choices for the places of the 3's in the sequence (recall the multinomial coefficient defined in (1.4.4)). Therefore, the probability function for the counts

(X_1, X_2, X_3) is equal to

$$\begin{aligned} p_{(X_1,X_2,X_3)}(x_1, x_2, x_3) &= \binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3}\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3} \\ &= \binom{n}{x_1\ x_2\ x_3}\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}. \end{aligned}$$

We say that

$$(X_1, X_2, X_3) \sim \text{ Multinomial}(n, \theta_1, \theta_2, \theta_3).$$

Notice that the Multinomial$(n, \theta_1, \theta_2, \theta_3)$ generalizes the Binomial(n, θ) distribution, as we are now counting the number of response values in three possible categories rather than two. Also, it is immediate that

$$X_i \sim \text{ Binomial}(n, \theta_i)$$

because X_i equals the number of occurrences of i in the n independent response values, and i occurs for an individual response with probability equal to θ_i (also see Problem 2.8.18).

As a simple example, suppose that we have an urn containing 10 red balls, 20 white balls, and 30 black balls. If we randomly draw 10 balls from the urn with replacement, what is the probability that we will obtain 3 red, 4 white, and 3 black balls? Because we are drawing with replacement, the draws are i.i.d., so the counts are distributed Multinomial$(10, 10/60, 20/60, 30/60)$. The required probability equals

$$\binom{10}{3\,4\,3}\left(\frac{10}{60}\right)^3\left(\frac{20}{60}\right)^4\left(\frac{30}{60}\right)^3 = 3.0007 \times 10^{-2}.$$

Note that if we had drawn without replacement, then the draws would not be i.i.d., the counts would thus not follow a multinomial distribution but rather a generalization of the hypergeometric distribution, as discussed in Problem 2.3.29.

Now suppose we have a response s that takes k possible values — for convenience, labelled $1, 2, \ldots, k$ — with the probability distribution given by $P(s = i) = \theta_i$. For a sample $(s_1, \ldots, s_n)$, define the counts $X_i = \sum_{j=1}^n I_{\{i\}}(s_j)$ for $i = 1, \ldots k$. Then, arguing as above and recalling the development of (1.4.4), we have

$$p_{(X_1,\ldots,X_k)}(x_1, \ldots, x_k) = \binom{n}{x_1\ \ldots\ x_k}\theta_1^{x_1}\cdots\theta_k^{x_k}$$

whenever each $x_i \in \{0, \ldots, n\}$ and $x_1 + \cdots + x_k = n$. In this case, we write

$$(X_1, \ldots, X_k) \sim \text{ Multinomial}(n, \theta_1, \ldots, \theta_k). \blacksquare$$

2.8.4 Order Statistics

Suppose now that $(X_1, \ldots, X_n)$ is a sample. In many applications of statistics, we will have n data values where the assumption that these arise as an i.i.d. sequence makes sense. It is often of interest, then, to order these from smallest to largest to obtain the *order statistics*

$$X_{(1)}, \ldots, X_{(n)}.$$

Here, $X_{(i)}$ is equal to the ith smallest value in the sample $X_1, \ldots, X_n$. So, for example, if $n = 5$ and

$$X_1 = 2.3,\ X_2 = 4.5,\ X_3 = -1.2,\ X_4 = 2.2,\ X_5 = 4.3$$

then

$$X_{(1)} = -1.2,\ X_{(2)} = 2.2,\ X_{(3)} = 2.3,\ X_{(4)} = 4.3,\ X_{(5)} = 4.5.$$

Of considerable interest in many situations are the distributions of the order statistics. Consider the following examples.

EXAMPLE 2.8.6 *Distribution of the Sample Maximum*
Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. so that $F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x)$. Then the *largest-order statistic* $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$ is the *maximum* of these n random variables.

Now $X_{(n)}$ is another random variable. What is its cumulative distribution function? We see that $X_{(n)} \le x$ if and only if $X_i \le x$ for all i. Hence,

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \le x) = P(X_1 \le x,\ X_2 \le x,\ \ldots, X_n \le x) \\ &= P(X_1 \le x)P(X_2 \le x)\cdots P(X_n \le x) = F_{X_1}(x)F_{X_2}(x)\cdots F_{X_n}(x) \\ &= \left(F_{X_1}(x)\right)^n. \end{aligned}$$

If F_{X_1} corresponds to an absolutely continuous distribution, then we can differentiate this expression to obtain the density of $X_{(n)}$. ▮

EXAMPLE 2.8.7
As a special case of Example 2.8.6, suppose that $X_1, X_2, \ldots, X_n$ are identically and independently distributed Uniform[0, 1]. From the above, for $0 \le x \le 1$, we have $F_{X_{(n)}}(x) = \left(F_{X_1}(x)\right)^n = x^n$. It then follows from Corollary 2.5.1 that the density $f_{X_{(n)}}$ of $X_{(n)}$ equals $f_{X_{(n)}}(x) = F'_{X_{(n)}}(x) = nx^{n-1}$ for $0 \le x \le 1$, with (of course) $f_{X_{(n)}}(x) = 0$ for $x < 0$ and $x > 1$. Note that, from Problem 2.4.24, we can write $X_{(n)} \sim \text{Beta}(n, 1)$. ▮

EXAMPLE 2.8.8 *Distribution of the Sample Minimum*
Following Example 2.8.6, we can also obtain the distribution function of the sample minimum, or *smallest-order statistic*, $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$. We have

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \le x) \\ &= 1 - P(X_{(1)} > x) \\ &= 1 - P(X_1 > x,\ X_2 > x,\ \ldots, X_n > x) \\ &= 1 - P(X_1 > x)\ P(X_2 > x)\cdots P(X_n > x) \\ &= 1 - \left(1 - F_{X_1}(x)\right)\left(1 - F_{X_2}(x)\right)\cdots\left(1 - F_{X_n}(x)\right) \\ &= 1 - \left(1 - F_{X_1}(x)\right)^n. \end{aligned}$$

Again, if F_{X_1} corresponds to an absolutely continuous distribution, we can differentiate this expression to obtain the density of $X_{(1)}$. ▮

EXAMPLE 2.8.9

Let $X_1, \ldots, X_n$ be i.i.d. Uniform[0, 1]. Hence, for $0 \le x \le 1$,

$$F_{X_{(1)}}(x) = P(X_{(1)} \le x) = 1 - P(X_{(1)} > x) = 1 - (1-x)^n.$$

It then follows from Corollary 2.5.1 that the density $f_{X_{(1)}}$ of $X_{(1)}$ satisfies $f_{X_{(1)}}(x) = F'_{X_{(1)}}(x) = n(1-x)^{n-1}$ for $0 \le x \le 1$, with (of course) $f_{X_{(1)}}(x) = 0$ for $x < 0$ and $x > 1$. Note that, from Problem 2.4.24, we can write $X_{(1)} \sim \text{Beta}(1, n)$. ▮

The sample median and sample quartiles are defined in terms of order statistics and used in statistical applications. These quantities, and their uses, are discussed in Section 5.5.

Summary of Section 2.8

- If X and Y are discrete, then the conditional probability function of Y, given X, equals $p_{Y|X}(y\,|\,x) = p_{X,Y}(x,y)/\,p_X(x)$.
- If X and Y are absolutely continuous, then the conditional density function of Y, given X, equals $f_{Y|X}(y\,|\,x) = f_{X,Y}(x,y)/f_X(x)$.
- X and Y are independent if $P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2)$ for all $B_1, B_2 \subseteq R^1$.
- Discrete X and Y are independent if and only if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all $x, y \in R^1$ or, equivalently, $p_{Y|X}(y\,|\,x) = p_Y(y)$.
- Absolutely continuous X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)\,f_Y(y)$ for all $x, y \in R^1$ or, equivalently, $f_{Y|X}(y\,|\,x) = f_Y(y)$.
- A sequence $X_1, X_2, \ldots, X_n$ is i.i.d. if the random variables are independent, and each X_i has the same distribution.

EXERCISES

2.8.1 Suppose X and Y have joint probability function

$$p_{X,Y}(x,y) = \begin{cases} 1/6 & x = -2\,,\ y = 3 \\ 1/12 & x = -2\,,\ y = 5 \\ 1/6 & x = 9\,,\ y = 3 \\ 1/12 & x = 9\,,\ y = 5 \\ 1/3 & x = 13\,,\ y = 3 \\ 1/6 & x = 13\,,\ y = 5 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $p_X(x)$ for all $x \in R^1$.
(b) Compute $p_Y(y)$ for all $y \in R^1$.
(c) Determine whether or not X and Y are independent.

2.8.2 Suppose X and Y have joint probability function

$$p_{X,Y}(x,y) = \begin{cases} 1/16 & x = -2\,,\ y = 3 \\ 1/4 & x = -2\,,\ y = 5 \\ 1/2 & x = 9\,,\ y = 3 \\ 1/16 & x = 9\,,\ y = 5 \\ 1/16 & x = 13\,,\ y = 3 \\ 1/16 & x = 13\,,\ y = 5 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $p_X(x)$ for all $x \in R^1$.
(b) Compute $p_Y(y)$ for all $y \in R^1$.
(c) Determine whether or not X and Y are independent.

2.8.3 Suppose X and Y have joint density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{12}{49}\left(2 + x + xy + 4y^2\right) & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $f_X(x)$ for all $x \in R^1$.
(b) Compute $f_Y(y)$ for all $y \in R^1$.
(c) Determine whether or not X and Y are independent.

2.8.4 Suppose X and Y have joint density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{2}{5(2+e)}\left(3 + e^x + 3y + 3ye^y + ye^x + ye^{x+y}\right) & \begin{array}{l} 0 \le x \le 1, \\ 0 \le y \le 1 \end{array} \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute $f_X(x)$ for all $x \in R^1$.
(b) Compute $f_Y(y)$ for all $y \in R^1$.
(c) Determine whether or not X and Y are independent.

2.8.5 Suppose X and Y have joint probability function

$$p_{X,Y}(x,y) = \begin{cases} 1/9 & x = -4,\ y = -2 \\ 2/9 & x = 5,\ y = -2 \\ 3/9 & x = 9,\ y = -2 \\ 2/9 & x = 9,\ y = 0 \\ 1/9 & x = 9,\ y = 4 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $P(Y = 4 \mid X = 9)$.
(b) Compute $P(Y = -2 \mid X = 9)$.
(c) Compute $P(Y = 0 \mid X = -4)$.
(d) Compute $P(Y = -2 \mid X = 5)$.
(e) Compute $P(X = 5 \mid Y = -2)$.

2.8.6 Let $X \sim$ Bernoulli(θ) and $Y \sim$ Geometric(θ), with X and Y independent. Let $Z = X + Y$. What is the probability function of Z?

2.8.7 For each of the following joint density functions $f_{X,Y}$ (taken from Exercise 2.7.4), compute the conditional density $f_{Y|X}(y \mid x)$, and determine whether or not X and Y are independent.

(a)

$$f_{X,Y}(x,y) = \begin{cases} 2x^2y + Cy^5 & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(b)

$$f_{X,Y}(x,y) = \begin{cases} C(xy + x^5y^5) & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(c)

$$f_{X,Y}(x,y) = \begin{cases} C(xy + x^5y^5) & 0 \le x \le 4,\ 0 \le y \le 10 \\ 0 & \text{otherwise.} \end{cases}$$

(d)

$$f_{X,Y}(x,y) = \begin{cases} Cx^5y^5 & 0 \le x \le 4,\ 0 \le y \le 10 \\ 0 & \text{otherwise.} \end{cases}$$

2.8.8 Let X and Y be jointly absolutely continuous random variables. Suppose $X \sim$ Exponential(2) and that $P(Y > 5 \mid X = x) = e^{-3x}$. Compute $P(Y > 5)$.

2.8.9 Give an example of two random variables X and Y, each taking values in the set $\{1, 2, 3\}$, such that $P(X = 1, Y = 1) = P(X = 1)\,P(Y = 1)$, but X and Y are *not* independent.

2.8.10 Let $X \sim$ Bernoulli(θ) and $Y \sim$ Bernoulli(ψ), where $0 < \theta < 1$ and $0 < \psi < 1$. Suppose $P(X = 1, Y = 1) = P(X = 1)\,P(Y = 1)$. Prove that X and Y must be independent.

2.8.11 Suppose that X is a *constant* random variable and that Y is any random variable. Prove that X and Y must be independent.

2.8.12 Suppose $X \sim$ Bernoulli$(1/3)$ and $Y \sim$ Poisson(λ), with X and Y independent and with $\lambda > 0$. Compute $P(X = 1 \mid Y = 5)$.

2.8.13 Suppose $P(X = x,\ Y = y) = 1/8$ for $x = 3, 5$ and $y = 1, 2, 4, 7$, otherwise $P(X = x,\ Y = y) = 0$.
(a) Compute the conditional probability function $p_{Y|X}(y|x)$ for all $x, y \in R^1$ with $p_X(x) > 0$.
(b) Compute the conditional probability function $p_{X|Y}(x|y)$ for all $x, y \in R^1$ with $p_Y(y) > 0$.
(c) Are X and Y independent? Why or why not?

2.8.14 Let X and Y have joint density $f_{X,Y}(x,y) = (x^2 + y)/36$ for $-2 < x < 1$ and $0 < y < 4$, otherwise $f_{X,Y}(x,y) = 0$.
(a) Compute the conditional density $f_{Y|X}(y|x)$ for all $x, y \in R^1$ with $f_X(x) > 0$.
(b) Compute the conditional density $f_{X|Y}(x|y)$ for all $x, y \in R^1$ with $f_Y(y) > 0$.
(c) Are X and Y independent? Why or why not?

2.8.15 Let X and Y have joint density $f_{X,Y}(x, y) = (x^2 + y)/4$ for $0 < x < y < 2$, otherwise $f_{X,Y}(x, y) = 0$. Compute each of the following.
(a) The conditional density $f_{Y|X}(y|x)$ for all $x, y \in R^1$ with $f_X(x) > 0$
(b) The conditional density $f_{X|Y}(x|y)$ for all $x, y \in R^1$ with $f_Y(y) > 0$
(c) Are X and Y independent? Why or why not?

2.8.16 Suppose we obtain the following sample of size $n = 6$: $X_1 = 12$, $X_2 = 8$, $X_3 = X_4 = 9$, $X_5 = 7$, and $X_6 = 11$. Specify the order statistics $X_{(i)}$ for $1 \le i \le 6$.

PROBLEMS

2.8.17 Let X and Y be jointly absolutely continuous random variables, having joint density of the form

$$f_{X,Y}(x, y) = \begin{cases} C_1(2x^2y + C_2y^5) & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Determine values of C_1 and C_2, such that $f_{X,Y}$ is a valid joint density function, and X and Y are independent.

2.8.18 Let X and Y be discrete random variables. Suppose $p_{X,Y}(x, y) = g(x)\,h(y)$, for some functions g and h. Prove that X and Y are independent. (Hint: Use Theorem 2.8.3(a) and Theorem 2.7.4.)

2.8.19 Let X and Y be jointly absolutely continuous random variables. Suppose $f_{X,Y}(x, y) = g(x)\,h(y)$, for some functions g and h. Prove that X and Y are independent. (Hint: Use Theorem 2.8.3(b) and Theorem 2.7.5.)

2.8.20 Let X and Y be discrete random variables, with $P(X = 1) > 0$ and $P(X = 2) > 0$. Suppose $P(Y = 1 \,|\, X = 1) = 3/4$ and $P(Y = 2 \,|\, X = 2) = 3/4$. Prove that X and Y cannot be independent.

2.8.21 Let X and Y have the bivariate normal distribution, as in Example 2.7.9. Prove that X and Y are independent if and only if $\rho = 0$.

2.8.22 Suppose that $(X_1, X_2, X_3) \sim \text{Multinomial}(n, \theta_1, \theta_2, \theta_3)$. Prove, by summing the joint probability function, that $X_1 \sim \text{Binomial}(n, \theta_1)$.

2.8.23 Suppose that $(X_1, X_2, X_3) \sim \text{Multinomial}(n, \theta_1, \theta_2, \theta_3)$. Find the conditional distribution of X_2 given that $X_1 = x_1$.

2.8.24 Suppose that $X_1, \ldots, X_n$ is a sample from the Exponential(λ) distribution. Find the densities $f_{X_{(1)}}$ and $f_{X_{(n)}}$.

2.8.25 Suppose that $X_1, \ldots, X_n$ is a sample from a distribution with cdf F. Prove that

$$F_{X_{(i)}}(x) = \sum_{j=i}^{n} \binom{n}{j} F^j(x)\,(1 - F(x))^{n-j}.$$

(Hint: Note that $X_{(i)} \le x$ if and only if at least i of $X_1, \ldots, X_n$ are less than or equal to x.)

2.8.26 Suppose that $X_1, \ldots, X_5$ is a sample from the Uniform[0, 1] distribution. If we define the sample median to be $X_{(3)}$, find the density of the sample median. Can you identify this distribution? (Hint: Use Problem 2.8.25.)

2.8.27 Suppose that $(X, Y) \sim$ Bivariate Normal$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Prove that Y given $X = x$ is distributed $N(\mu_2 + \rho\sigma_2 (x - \mu_1)/\sigma_1, \left(1 - \rho^2\right)\sigma_2^2)$. Establish the analogous result for the conditional distribution of X given $Y = y$. (Hint: Use (2.7.1) for Y given $X = x$ and its analog for X given $Y = y$.)

CHALLENGES

2.8.28 Let X and Y be random variables.
(a) Suppose X and Y are both discrete. Prove that X and Y are independent if and only if $P(Y = y \,|\, X = x) = P(Y = y)$ for all x and y such that $P(X = x) > 0$.
(b) Suppose X and Y are jointly absolutely continuous. Prove that X and Y are independent if and only if $P(a \le Y \le b \,|\, X = x) = P(a \le Y \le b)$ for all x and y such that $f_X(x) > 0$.

2.9 Multidimensional Change of Variable

Let X and Y be random variables with known joint distribution. Suppose that $Z = h_1(X, Y)$ and $W = h_2(X, Y)$, where $h_1, h_2 : R^2 \to R^1$ are two functions. What is the joint distribution of Z and W?

This is similar to the problem considered in Section 2.6, except that we have moved from a one-dimensional to a two-dimensional setting. The two-dimensional setting is more complicated; however, the results remain essentially the same, as we shall see.

2.9.1 The Discrete Case

If X and Y are *discrete* random variables, then the distribution of Z and W is essentially straightforward.

Theorem 2.9.1 Let X and Y be discrete random variables, with joint probability function $p_{X,Y}$. Let $Z = h_1(X, Y)$ and $W = h_2(X, Y)$, where $h_1, h_2 : R^2 \to R^1$ are some functions. Then Z and W are also discrete, and their joint probability function $p_{Z,W}$ satisfies

$$p_{Z,W}(z, w) = \sum_{\substack{x,y \\ h_1(x,y)=z,\, h_2(x,y)=w}} p_{X,Y}(x, y).$$

Here, the sum is taken over all pairs (x, y) such that $h_1(x, y) = z$ and $h_2(x, y) = w$.

PROOF We compute that $p_{Z,W}(z, w) = P(Z = z, W = w) = P(h_1(X, Y) = z, h_2(X, Y) = w)$. This equals

$$\sum_{\substack{x,y \\ h_1(x,y)=z,\, h_2(x,y)=w}} P(X = x, Y = y) = \sum_{\substack{x,y \\ h_1(x,y)=z,\, h_2(x,y)=w}} p_{X,Y}(x, y),$$

as claimed. ▮

As a special case, we note the following.

Corollary 2.9.1 Suppose in the context of Theorem 2.9.1 that the joint function $h = (h_1, h_2) : R^2 \to R^2$ defined by $h(x, y) = \big(h_1(x, y)\,,\; h_2(x, y)\big)$ is one-to-one, i.e., if $h_1(x_1, y_1) = h_1(x_2, y_2)$ and $h_2(x_1, y_1) = h_2(x_2, y_2)$, then $x_1 = x_2$ and $y_1 = y_2$. Then

$$p_{Z,W}(z, w) = p_{X,Y}(h^{-1}(z, w)),$$

where $h^{-1}(z, w)$ is the unique pair (x, y) such that $h(x, y) = (z, w)$.

EXAMPLE 2.9.1

Suppose X and Y have joint density function

$$p_{X,Y}(x, y) = \begin{cases} 1/6 & x = 2, y = 6 \\ 1/12 & x = -2, y = -6 \\ 1/4 & x = -3, y = 11 \\ 1/2 & x = 3, y = -8 \\ 0 & \text{otherwise.} \end{cases}$$

Let $Z = X + Y$ and $W = Y - X^2$. Then $p_{Z,W}(8, 2) = P(Z = 8, W = 2) = P(X = 2, Y = 6) + P(X = -3, Y = 11) = 1/6 + 1/4 = 5/12$. On the other hand, $p_{Z,W}(-5, -17) = P(Z = -5, W = -17) = P(X = 3, Y = -8) = \frac{1}{2}$. ∎

2.9.2 The Continuous Case (Advanced)

If X and Y are *continuous*, and the function $h = (h_1, h_2)$ is *one-to-one*, then it is again possible to compute a formula for the joint density of Z and W, as the following theorem shows. To state it, recall from multivariable calculus that, if $h = (h_1, h_2) : R^2 \to R^2$ is a differentiable function, then its *Jacobian derivative* J is defined by

$$J(x, y) = \det \begin{pmatrix} \frac{\partial h_1}{\partial x} & \frac{\partial h_2}{\partial x} \\ \frac{\partial h_1}{\partial y} & \frac{\partial h_2}{\partial y} \end{pmatrix} = \frac{\partial h_1}{\partial x}\frac{\partial h_2}{\partial y} - \frac{\partial h_2}{\partial x}\frac{\partial h_1}{\partial y}.$$

Theorem 2.9.2 Let X and Y be jointly absolutely continuous, with joint density function $f_{X,Y}$. Let $Z = h_1(X, Y)$ and $W = h_2(X, Y)$, where $h_1, h_2 : R^2 \to R^1$ are differentiable functions. Define the joint function $h = (h_1, h_2) : R^2 \to R^2$ by

$$h(x, y) = (h_1(x, y), h_2(x, y))\,.$$

Assume that h is one-to-one, at least on the region $\{(x, y) : f(x, y) > 0\}$, i.e., if $h_1(x_1, y_1) = h_1(x_2, y_2)$ and $h_2(x_1, y_1) = h_2(x_2, y_2)$, then $x_1 = x_2$ and $y_1 = y_2$. Then Z and W are also jointly absolutely continuous, with joint density function $f_{Z,W}$ given by

$$f_{Z,W}(z, w) = f_{X,Y}(h^{-1}(z, w)) \,/\, |J(h^{-1}(z, w))|,$$

where J is the Jacobian derivative of h and where $h^{-1}(z, w)$ is the unique pair (x, y) such that $h(x, y) = (z, w)$.

PROOF See Section 2.11 for the proof of this result. ▮

EXAMPLE 2.9.2

Let X and Y be jointly absolutely continuous, with joint density function $f_{X,Y}$ given by

$$f_{X,Y}(x,y) = \begin{cases} 4x^2y + 2y^5 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

as in Example 2.7.6. Let $Z = X + Y^2$ and $W = X - Y^2$. What is the joint density of Z and W?

We first note that $Z = h_1(X,Y)$ and $W = h_2(X,Y)$, where $h_1(x,y) = x + y^2$ and $h_2(x,y) = x - y^2$. Hence,

$$J(x,y) = \frac{\partial h_1}{\partial x}\frac{\partial h_2}{\partial y} - \frac{\partial h_2}{\partial x}\frac{\partial h_1}{\partial y} = (1)(-2y) - (1)(2y) = -4y.$$

We may *invert* the relationship h by solving for X and Y, to obtain that

$$X = \frac{1}{2}(Z + W) \text{ and } Y = \sqrt{\frac{Z - W}{2}}.$$

This means that $h = (h_1, h_2)$ is invertible, with

$$h^{-1}(z,w) = \left(\frac{1}{2}(z+w)\,,\ \sqrt{\frac{z-w}{2}}\right).$$

Hence, using Theorem 2.9.2, we see that

$$\begin{aligned}
& f_{Z,W}(z,w) \\
= \;& f_{X,Y}(h^{-1}(z,w)) \,/\, |J(h^{-1}(z,w))| \\
= \;& f_{X,Y}\left(\frac{1}{2}(z+w),\ \sqrt{\frac{z-w}{2}}\right) / \,|J(h^{-1}(z,w))| \\
= \;& \begin{cases} \left\{4(\tfrac{1}{2}(z+w))^2\sqrt{\tfrac{z-w}{2}} + 2\left(\sqrt{\tfrac{z-w}{2}}\right)^5\right\} /4\sqrt{\tfrac{z-w}{2}} & 0 \le \tfrac{1}{2}(z+w) \le 1, \\ & 0 \le \sqrt{\tfrac{z-w}{2}} \le 1 \\ 0 & \text{otherwise} \end{cases} \\
= \;& \begin{cases} (\tfrac{z+w}{2})^2 + \tfrac{1}{2}\left(\tfrac{z-w}{2}\right)^2 & 0 \le z+w \le 2,\ 0 \le z-w \le 2 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

We have thus obtained the joint density function for Z and W. ▮

EXAMPLE 2.9.3

Let U_1 and U_2 be independent, each having the Uniform[0, 1] distribution. (We could write this as U_1, U_2 are i.i.d. Uniform[0, 1].) Thus,

$$f_{U_1,U_2}(u_1,u_2) = \begin{cases} 1 & 0 \le u_1 \le 1,\ 0 \le u_2 \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then define X and Y by

$$X = \sqrt{2\,\log(1/U_1)}\,\cos(2\pi U_2), \quad Y = \sqrt{2\,\log(1/U_1)}\,\sin(2\pi U_2).$$

What is the joint density of X and Y?

We see that here $X = h_1(U_1, U_2)$ and $Y = h_2(U_1, U_2)$, where

$$h_1(u_1, u_2) = \sqrt{2\,\log(1/u_1)}\,\cos(2\pi u_2), \quad h_2(u_1, u_2) = \sqrt{2\,\log(1/u_1)}\,\sin(2\pi u_2).$$

Therefore,

$$\frac{\partial h_1}{\partial u_1}(u_1, u_2) = \frac{1}{2}\,(2\,\log(1/u_1))^{-1/2}\left(2u_1(-1/u_1^2)\right)\,\cos(2\pi u_2).$$

Continuing in this way, we eventually compute (see Exercise 2.9.1) that

$$J(u_1, u_2) = \frac{\partial h_1}{\partial u_1}\frac{\partial h_2}{\partial u_2} - \frac{\partial h_2}{\partial u_1}\frac{\partial h_1}{\partial u_2} = -\frac{2\pi}{u_1}\left(\cos^2(2\pi u_2) + \sin^2(2\pi u_2)\right) = -\frac{2\pi}{u_1}.$$

On the other hand, inverting the relationship h, we compute that

$$U_1 = e^{-(X^2+Y^2)/2}, \quad U_2 = \arctan(Y/X)\,/\,2\pi\,.$$

Hence, using Theorem 2.9.2, we see that

$$\begin{aligned}
f_{X,Y}(x, y) &= f_{U_1,U_2}(h^{-1}(x, y))\,/\,|J(h^{-1}(x, y))| \\
&= f_{U_1,U_2}\left(e^{-(x^2+y^2)/2},\ \arctan(y/x)\,/2\pi\right) \\
&\quad \times \left|J\left(e^{-(x^2+y^2)/2},\ \arctan(y/x)\,/2\pi\right)\right|^{-1} \\
&= \begin{cases} 1\,/\,|-2\pi\ /\,e^{-(x^2+y^2)/2}| & \begin{array}{l} 0 \le e^{-(x^2+y^2)/2} \le 1, \\ 0 \le \arctan(y/x)\,/2\pi\ \le 1 \end{array} \\ 0 & \text{otherwise} \end{cases} \\
&= \frac{1}{2\pi}\,e^{-(x^2+y^2)/2}
\end{aligned}$$

where the last expression is valid for all x and y, because we *always* have

$$0 \le e^{-(x^2+y^2)/2} \le 1$$

and $0 \le \arctan(y/x)\,/2\pi\ \le 1$.

We conclude that

$$f_{X,Y}(x, y) = \left(\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\right).$$

We recognize this as a product of two standard normal densities. We thus conclude that $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ and that, furthermore, X and Y are independent. ∎

2.9.3 Convolution

Suppose now that X and Y are independent, with known distributions, and that $Z = X + Y$. What is the distribution of Z? In this case, the distribution of Z is called the *convolution* of the distributions of X and of Y. Fortunately, the convolution is often reasonably straightforward to compute.

Theorem 2.9.3 Let X and Y be independent, and let $Z = X + Y$.
(a) If X and Y are both discrete, with probability functions p_X and p_Y, then Z is also discrete, with probability function p_Z given by

$$p_Z(z) = \sum_w p_X(z - w)p_Y(w).$$

(b) If X and Y are jointly absolutely continuous, with density functions f_X and f_Y, then Z is also absolutely continuous, with density function f_Z given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - w)\, f_Y(w)\, dw.$$

PROOF (a) We let $W = Y$ and consider the two-dimensional transformation from (X, Y) to $(Z, W) = (X + Y, Y)$.

In the discrete case, by Corollary 2.9.1, $p_{Z,W}(z, w) = p_{X,Y}(z - w, w)$. Then from Theorem 2.7.4, $p_Z(z) = \sum_w p_{Z,W}(z, w) = \sum_w p_{X,Y}(z - w, w)$. But because X and Y are independent, $p_{X,Y}(x, y) = p_X(x)\, p_Y(y)$, so $p_{X,Y}(z - w, w) = p_X(z - w)\, p_Y(w)$. This proves part (a).

(b) In the continuous case, we must compute the Jacobian derivative $J(x, y)$ of the transformation from (X, Y) to $(Z, W) = (X + Y, Y)$. Fortunately, this is very easy, as we obtain

$$J(x, y) = \frac{\partial(x + y)}{\partial x}\frac{\partial y}{\partial y} - \frac{\partial y}{\partial x}\frac{\partial(x + y)}{\partial y} = (1)(1) - (0)(1) = 1.$$

Hence, from Theorem 2.9.2, $f_{Z,W}(z, w) = f_{X,Y}(z - w, w)/|1| = f_{X,Y}(z - w, w)$ and from Theorem 2.7.5,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z,W}(z, w)\, dw = \int_{-\infty}^{\infty} f_{X,Y}(z - w, w)\, dw.$$

But because X and Y are independent, we may take $f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$, so $f_{X,Y}(z - w, w) = f_X(z - w)\, f_Y(w)$. This proves part (b). ▮

EXAMPLE 2.9.4

Let $X \sim$ Binomial(4, 1/5) and $Y \sim$ Bernoulli(1/4), with X and Y independent. Let $Z = X + Y$. Then

$$\begin{aligned} p_Z(3) &= P(X + Y = 3) = P(X = 3, Y = 0) + P(X = 2, Y = 1) \\ &= \binom{4}{3}(1/5)^3(4/5)^1\,(3/4) + \binom{4}{2}(1/5)^2(4/5)^2\,(1/4) \\ &= 4(1/5)^3(4/5)^1\,(3/4) + 6(1/5)^2(4/5)^2\,(1/4) \doteq 0.0576. \end{aligned}$$ ▮

EXAMPLE 2.9.5
Let $X \sim$ Uniform$[3, 7]$ and $Y \sim$ Exponential(6), with X and Y independent. Let $Z = X + Y$. Then

$$\begin{aligned} f_Z(5) &= \int_{-\infty}^{\infty} f_X(x)\, f_Y(5-x)\, dx = \int_3^5 (1/4)\, 6\, e^{-6(5-x)}\, dx \\ &= -(1/4)e^{-6(5-x)} \Big|_{x=3}^{x=5} = -(1/4)e^{-12} + (1/4)e^0 \doteq 0.2499985. \end{aligned}$$

Note that here the limits of integration go from 3 to 5 only, because $f_X(x) = 0$ for $x < 3$, while $f_Y(5-x) = 0$ for $x > 5$. ∎

Summary of Section 2.9

- If X and Y are discrete, and $Z = h_1(X, Y)$ and $W = h_2(X, Y)$, then

$$p_{Z,W}(z, w) = \sum_{\{(x,y):\ h_1(x,y)=z,\ h_2(x,y)=w\}} p_{X,Y}(x, y).$$

- If X and Y are absolutely continuous, if $Z = h_1(X, Y)$ and $W = h_2(X, Y)$, and if $h = (h_1, h_2) : R^2 \to R^2$ is one-to-one with Jacobian $J(x, y)$, then $f_{Z,W}(z, w) = f_{X,Y}(h^{-1}(z, w))/|J(h^{-1}(z, w))|$.
- This allows us to compute the joint distribution of functions of pairs of random variables.

EXERCISES

2.9.1 Verify explicitly in Example 2.9.3 that $J(u_1, u_2) = -2\pi/u_1$.

2.9.2 Let $X \sim$ Exponential(3) and $Y \sim$ Uniform$[1, 4]$, with X and Y independent. Let $Z = X + Y$ and $W = X - Y$.
(a) Write down the joint density $f_{X,Y}(x, y)$ of X and Y. (Be sure to consider the ranges of valid x and y values.)
(b) Find a two-dimensional function h such that $(Z, W) = h(X, Y)$.
(c) Find a two-dimensional function h^{-1} such that $(X, Y) = h^{-1}(Z, W)$.
(d) Compute the joint density $f_{Z,W}(z, w)$ of Z and W. (Again, be sure to consider the ranges of valid z and w values.)

2.9.3 Repeat parts (b) through (d) of Exercise 2.9.2, for the same random variables X and Y, if instead $Z = X^2 + Y^2$ and $W = X^2 - Y^2$.

2.9.4 Repeat parts (b) through (d) of Exercise 2.9.2, for the same random variables X and Y, if instead $Z = X + 4$ and $W = Y - 3$.

2.9.5 Repeat parts (b) through (d) of Exercise 2.9.2, for the same random variables X and Y, if instead $Z = Y^4$ and $W = X^4$.

2.9.6 Suppose the joint probability function of X and Y is given by

$$p_{X,Y}(x,y) = \begin{cases} 1/7 & x = 5, y = 0 \\ 1/7 & x = 5, y = 3 \\ 1/7 & x = 5, y = 4 \\ 3/7 & x = 8, y = 0 \\ 1/7 & x = 8, y = 4 \\ 0 & \text{otherwise.} \end{cases}$$

Let $Z = X + Y$, $W = X - Y$, $A = X^2 + Y^2$, and $B = 2X - 3Y^2$.
(a) Compute the joint probability function $p_{Z,W}(z, w)$.
(b) Compute the joint probability function $p_{A,B}(a, b)$.
(c) Compute the joint probability function $p_{Z,A}(z, a)$.
(d) Compute the joint probability function $p_{W,B}(w, b)$.

2.9.7 Let X have probability function

$$p_X(x) = \begin{cases} 1/3 & x = 0 \\ 1/2 & x = 2 \\ 1/6 & x = 3 \\ 0 & \text{otherwise,} \end{cases}$$

and let Y have probability function

$$p_Y(y) = \begin{cases} 1/6 & y = 2 \\ 1/12 & y = 5 \\ 3/4 & y = 9 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose X and Y are independent. Let $Z = X + Y$. Compute $p_Z(z)$ for all $z \in R^1$.

2.9.8 Let $X \sim \text{Geometric}(1/4)$, and let Y have probability function

$$p_Y(y) = \begin{cases} 1/6 & y = 2 \\ 1/12 & y = 5 \\ 3/4 & y = 9 \\ 0 & \text{otherwise.} \end{cases}$$

Let $W = X + Y$. Suppose X and Y are independent. Compute $p_W(w)$ for all $w \in R^1$.

2.9.9 Suppose X and Y are discrete, with $P(X = 1, Y = 1) = P(X = 1, Y = 2) = P(X = 1, Y = 3) = P(X = 2, Y = 2) = P(X = 2, Y = 3) = 1/5$, otherwise $P(X = x, Y = y) = 0$. Let $Z = X - Y^2$ and $W = X^2 + 5Y$.
(a) Compute the joint probability function $p_{Z,W}(z, w)$ for all $z, w \in R^1$.
(b) Compute the marginal probability function $p_Z(z)$ for Z.
(c) Compute the marginal probability function $p_W(w)$ for W.

2.9.10 Suppose X has density $f_X(x) = x^3/4$ for $0 < x < 2$, otherwise $f_X(x) = 0$, and Y has density $f_Y(y) = 5y^4/32$ for $0 < y < 2$, otherwise $f_Y(y) = 0$. Assume X and Y are independent, and let $Z = X + Y$.

(a) Compute the joint density $f_{X,Y}(x, y)$ for all $x, y \in R^1$.
(b) Compute the density $f_Z(z)$ for Z.

PROBLEMS

2.9.11 Suppose again that X has density $f_X(x) = x^3/4$ for $0 < x < 2$, otherwise $f_X(x) = 0$, that Y has density $f_Y(y) = 5y^4/32$ for $0 < y < 2$, otherwise $f_Y(y) = 0$, and that X and Y are independent. Let $Z = X - Y$ and $W = 4X + 3Y$.
(a) Compute the joint density $f_{Z,W}(z, w)$ for all $z, w \in R^1$.
(b) Compute the marginal density $f_Z(z)$ for Z.
(c) Compute the marginal density $f_W(w)$ for W.

2.9.12 Let $X \sim \text{Binomial}(n_1, \theta)$ independent of $Y \sim \text{Binomial}(n_2, \theta)$. Let $Z = X + Y$. Use Theorem 2.9.3(a) to prove that $Z \sim \text{Binomial}(n_1 + n_2, \theta)$.

2.9.13 Let X and Y be independent, with $X \sim \text{Negative-Binomial}(r_1, \theta)$ and $Y \sim \text{Negative-Binomial}(r_2, \theta)$. Let $Z = X + Y$. Use Theorem 2.9.3(a) to prove that $Z \sim \text{Negative-Binomial}(r_1 + r_2, \theta)$.

2.9.14 Let X and Y be independent, with $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. Let $Z = X + Y$. Use Theorem 2.9.3(b) to prove that $Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2.9.15 Let X and Y be independent, with $X \sim \text{Gamma}(\alpha_1, \lambda)$ and $Y \sim \text{Gamma}(\alpha_2, \lambda)$. Let $Z = X + Y$. Use Theorem 2.9.3(b) to prove that $Z \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda)$.

2.9.16 (MV) Show that when Z_1, Z_2 are i.i.d. $N(0, 1)$ and X, Y are given by (2.7.1), then $(X, Y) \sim \text{Bivariate Normal}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

2.10 Simulating Probability Distributions

So far, we have been concerned primarily with mathematical theory and manipulations of probabilities and random variables. However, modern high-speed computers can be used to simulate probabilities and random variables *numerically*. Such simulations have many applications, including:

- To approximate quantities that are too difficult to compute mathematically
- To graphically simulate complicated physical or biological systems
- To randomly sample from large data sets to search for errors or illegal activities, etc.
- To implement complicated algorithms to sharpen pictures, recognize speech, etc.
- To simulate intelligent behavior
- To encrypt data or generate passwords
- To solve puzzles or break codes by trying lots of random solutions
- To generate random choices for online quizzes, computer games, etc.

Indeed, as computers become faster and more widespread, probabilistic simulations are becoming more and more common in software applications, scientific research, quality control, marketing, law enforcement, etc.

In most applications of probabilistic simulation, the first step is to simulate random variables having certain distributions. That is, a certain probability distribution will be specified, and we want to generate one or more random variables having that distribution.

Now, nearly all modern computer languages come with a *pseudorandom number generator*, which is a device for generating a sequence $U_1, U_2, \ldots$ of random values that are approximately independent and have approximately the uniform distribution on $[0, 1]$. Now, in fact, the U_i are usually generated from some sort of deterministic iterative procedure, which is designed to "appear" random. So the U_i are, in fact, not random, but rather *pseudorandom*.

Nevertheless, we shall ignore any concerns about pseudorandomness and shall simply assume that

$$U_1, U_2, U_3, \ldots \sim \text{Uniform}[0, 1], \tag{2.10.1}$$

i.e., the U_i are i.i.d. Uniform$[0, 1]$.

Hence, if all we ever need are Uniform$[0, 1]$ random variables, then according to (2.10.1), we are all set. However, in most applications, other kinds of randomness are also required. We therefore consider how to use the uniform random variables of (2.10.1) to generate random variables having other distributions.

EXAMPLE 2.10.1 *The Uniform$[L, R]$ Distribution*
Suppose we want to generate $X \sim \text{Uniform}[L, R]$. According to Exercise 2.6.1, we can simply set

$$X = (R - L)U_1 + L,$$

to ensure that $X \sim \text{Uniform}[L, R]$. ▮

2.10.1 Simulating Discrete Distributions

We now consider the question of how to simulate from discrete distributions.

EXAMPLE 2.10.2 *The Bernoulli(θ) Distribution*
Suppose we want to generate $X \sim \text{Bernoulli}(\theta)$, where $0 < \theta < 1$. We can simply set

$$X = \begin{cases} 1 & U_1 \le \theta \\ 0 & U_1 > \theta. \end{cases}$$

Then clearly, we always have either $X = 0$ or $X = 1$. Furthermore, $P(X = 1) = P(U_1 \le \theta) = \theta$, because $U_1 \sim \text{Uniform}[0, 1]$. Hence, we see that $X \sim \text{Bernoulli}(\theta)$. ▮

EXAMPLE 2.10.3 *The Binomial(n, θ) Distribution*
Suppose we want to generate $Y \sim \text{Binomial}(n, \theta)$, where $0 < \theta < 1$ and $n \ge 1$. There are two natural methods for doing this.

First, we can simply define Y as follows:

$$Y = \min\{j : \sum_{k=0}^{j} \binom{n}{k} \theta^k (1-\theta)^{n-k} \geq U_1\}.$$

That is, we let Y be the largest value of j such that the sum of the binomial probabilities up to $j-1$ is still no more than U_1. In that case,

$$\begin{aligned} P(Y = y) &= P\left(\begin{array}{c} \sum_{k=0}^{y-1} \binom{n}{k} \theta^k (1-\theta)^{n-k} < U_1 \\ \text{and } \sum_{k=0}^{y} \binom{n}{k} \theta^k (1-\theta)^{n-k} \geq U_1 \end{array} \right) \\ &= P\left(\sum_{k=0}^{y-1} \binom{n}{k} \theta^k (1-\theta)^{n-k} < U_1 \leq \sum_{k=0}^{y} \binom{n}{k} \theta^k (1-\theta)^{n-k} \right) \\ &= \sum_{k=0}^{y} \binom{n}{k} \theta^k (1-\theta)^{n-k} - \sum_{k=0}^{y-1} \binom{n}{k} \theta^k (1-\theta)^{n-k} \\ &= \binom{n}{y} \theta^y (1-\theta)^{n-y}. \end{aligned}$$

Hence, we have $Y \sim \text{Binomial}(n, \theta)$, as desired.

Alternatively, we can set

$$X_i = \begin{cases} 1 & U_i \leq \theta \\ 0 & U_i > \theta \end{cases}$$

for $i = 1, 2, 3, \ldots$. Then, by Example 2.10.2, we have $X_i \sim \text{Bernoulli}(\theta)$ for each i, with the $\{X_i\}$ independent because the $\{U_i\}$ are independent. Hence, by the observation at the end of Example 2.3.3, if we set $Y = X_1 + \cdots + X_n$, then we will again have $Y \sim \text{Binomial}(n, \theta)$. ▮

In Example 2.10.3, the second method is more elegant and is also simpler computationally (as it does not require computing any binomial coefficients). On the other hand, the first method of Example 2.10.3 is more general, as the following theorem shows.

Theorem 2.10.1 Let p be a probability function for a discrete probability distribution. Let $x_1 < x_2 < x_3 < \cdots$ be all the values for which $p(x_i) > 0$. Let $U_1 \sim \text{Uniform}[0, 1]$. Define Y by

$$Y = \min\{x_j : \sum_{k=1}^{j} p(x_k) \geq U_1\}.$$

Then Y is a discrete random variable, having probability function p.

PROOF We have

$$\begin{aligned}
P(Y = x_i) &= P\left(\sum_{k=1}^{i-1} p(x_k) < U_1, \text{ and } \sum_{k=1}^{i} p(x_k) \geq U_1\right) \\
&= P\left(\sum_{k=1}^{i-1} p(x_k) < U_1 \leq \sum_{k=1}^{i} p(x_k)\right) \\
&= \sum_{k=1}^{i} p(x_k) - \sum_{k=1}^{i-1} p(x_k) = p(x_i).
\end{aligned}$$

Also, clearly $P(Y = y) = 0$ if $y \notin \{x_1, x_2, \ldots\}$. Hence, for all $y \in R^1$, we have $P(Y = y) = p(y)$, as desired. ▮

EXAMPLE 2.10.4 *The Geometric(θ) Distribution*
To simulate $Y \sim$ Geometric(θ), we again have two choices. Using Theorem 2.10.1, we can let $U_1 \sim$ Uniform[0, 1] and then set

$$\begin{aligned}
Y &= \min\{j : \sum_{k=0}^{j} \theta(1-\theta)^k \geq U_1\} = \min\{j : 1 - (1-\theta)^{j+1} \geq U_1\} \\
&= \min\{j : j \geq \frac{\log(1-U_1)}{\log(1-\theta)} - 1\} = \left\lfloor \frac{\log(1-U_1)}{\log(1-\theta)} \right\rfloor,
\end{aligned}$$

where $\lfloor r \rfloor$ means to round down r to the next integer value, i.e., $\lfloor r \rfloor$ is the *greatest integer* not exceeding r (sometimes called the *floor* of r).

Alternatively, using the definition of Geometric(θ) from Example 2.3.4, we can set

$$X_i = \begin{cases} 1 & U_i \leq \theta \\ 0 & U_i > \theta \end{cases}$$

for $i = 1, 2, 3, \ldots$ (where $U_i \sim$ Uniform[0, 1]), and then let $Y = \min\{i : X_i = 1\}$. Either way, we have $Y \sim$ Geometric(θ), as desired. ▮

2.10.2 Simulating Continuous Distributions

We next turn to the subject of simulating absolutely continuous distributions. In general, this is not an easy problem. However, for certain particular continuous distributions, it is not difficult, as we now demonstrate.

EXAMPLE 2.10.5 *The Uniform[L, R] Distribution*
We have already seen in Example 2.10.1 that if $U_1 \sim$ Uniform[0, 1], and we set

$$X = (R - L)U_1 + L,$$

then $X \sim$ Uniform[L, R]. Thus, simulating from any uniform distribution is straightforward. ▮

EXAMPLE 2.10.6 *The Exponential*(λ) *Distribution*
We have also seen, in Example 2.6.6, that if $U_1 \sim$ Uniform[0, 1], and we set

$$Y = \ln(1/U_1),$$

then $Y \sim$ Exponential(1). Thus, simulating from the Exponential(1) distribution is straightforward.

Furthermore, we know from Exercise 2.6.4 that once $Y \sim$ Exponential(1), then if $\lambda > 0$, and we set

$$Z = Y \,/\, \lambda = \ln(1/U_1)/\lambda,$$

then $Z \sim$ Exponential(λ). Thus, simulating from any Exponential(λ) distribution is also straightforward. ▮

EXAMPLE 2.10.7 *The* $N(\mu, \sigma^2)$ *Distribution*
Simulating from the standard normal distribution, $N(0, 1)$, may appear to be more difficult. However, by Example 2.9.3, if $U_1 \sim$ Uniform[0, 1] and $U_2 \sim$ Uniform[0, 1], with U_1 and U_2 independent, and we set

$$X = \sqrt{2\,\log(1/U_1)}\,\cos(2\pi U_2), \quad Y = \sqrt{2\,\log(1/U_1)}\,\sin(2\pi U_2), \tag{2.10.2}$$

then $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ (and furthermore, X and Y are independent). So, using this trick, the standard normal distribution can be easily simulated as well.

It then follows from Exercise 2.6.3 that, once we have $X \sim N(0, 1)$, if we set $Z = \sigma X + \mu$, then $Z \sim N(\mu, \sigma^2)$. Hence, it is straightforward to sample from any normal distribution. ▮

These examples illustrate that, for certain special continuous distributions, sampling from them is straightforward. To provide a *general* method of sampling from a continuous distribution, we first state the following definition.

Definition 2.10.1 Let X be a random variable, with cumulative distribution function F. Then the *inverse cdf* (or *quantile function*) of X is the function F^{-1} defined by

$$F^{-1}(t) = \min\{x : F(x) \geq t\},$$

for $0 < t < 1$.

In Figure 2.10.1, we have provided a plot of the inverse cdf of an $N(0, 1)$ distribution. Note that this function goes to $-\infty$ as the argument goes to 0, and goes to ∞ as the argument goes to 1.

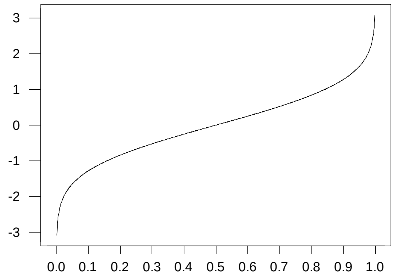

Figure 2.10.1: The inverse cdf of the $N(0, 1)$ distribution.

Using the inverse cdf, we obtain a general method of sampling from a continuous distribution, as follows.

Theorem 2.10.2 (*Inversion method for generating random variables*) Let F be any cumulative distribution function, and let $U \sim$ Uniform[0, 1]. Define a random variable Y by $Y = F^{-1}(U)$. Then $P(Y \le y) = F(y)$, i.e., Y has cumulative distribution function given by F.

PROOF We begin by noting that $P(Y \le y) = P(F^{-1}(U) \le y)$. But $F^{-1}(U)$ is the smallest value x such that $F(x) \ge U$. Hence, $F^{-1}(U) \le y$ if and only if $F(y) \ge U$, i.e., $U \le F(y)$. Therefore,

$$P(Y \le y) = P(F^{-1}(U) \le y) = P(U \le F(y)).$$

But $0 \le F(y) \le 1$, and $U \sim$ Uniform[0, 1], so $P(U \le F(y)) = F(y)$. Thus,

$$P(Y \le y) = P(U \le F(y)) = F(y).$$

It follows that F is the cdf of Y, as claimed. ▮

We note that Theorem 2.10.2 is valid for *any* cumulative distribution function, whether it corresponds to a continuous distribution, a discrete distribution, or a mixture of the two (as in Section 2.5.4). In fact, this was proved for discrete distributions in Theorem 2.10.1.

EXAMPLE 2.10.8 *Generating from an Exponential Distribution*
Let F be the cdf of an Exponential(1) random variable. Then

$$F(x) = \int_0^x e^{-t}\, dt = 1 - e^{-x}.$$

It then follows that

$$\begin{aligned} F^{-1}(t) &= \min\{x : F(x) \ge t\} = \min\{x\ : 1 - e^{-x} \ge t\} \\ &= \min\{x : x \ge -\ln(1-t)\} = -\ln(1-t) = \ln(1/(1-t)). \end{aligned}$$

Therefore, by Theorem 2.10.2, if $U \sim$ Uniform[0, 1], and we set

$$Y = F^{-1}(U) = \ln(1/(1-U)), \tag{2.10.3}$$

then $Y \sim \text{Exponential}(1)$.

Now, we have already seen from Example 2.6.6 that, if $U \sim$ Uniform[0, 1], and we set $Y = \ln(1/U)$, then $Y \sim \text{Exponential}(1)$. This is essentially the same as (2.10.3), except that we have replaced U by $1 - U$. On the other hand, this is not surprising, because we already know by Exercise 2.6.2 that, if $U \sim$ Uniform[0, 1], then also $1 - U \sim$ Uniform[0, 1]. ∎

EXAMPLE 2.10.9 *Generating from the Standard Normal Distribution*
Let Φ be the cdf of a $N(0, 1)$ random variable, as in Definition 2.5.2. Then

$$\Phi^{-1}(t) = \min\{x : \Phi(x) \geq t\},$$

and there is no simpler formula for $\Phi^{-1}(t)$. By Theorem 2.10.2, if
$U \sim$ Uniform[0, 1], and we set

$$Y = \Phi^{-1}(U), \tag{2.10.4}$$

then $Y \sim N(0, 1)$.

On the other hand, due to the difficulties of computing with Φ and Φ^{-1}, the method of (2.10.4) is not very practical. It is far better to use the method of (2.10.2), to simulate a normal random variable. ∎

For distributions that are too complicated to sample using the inversion method of Theorem 2.10.2, and for which no simple trick is available, it may still be possible to do sampling using *Markov chain methods*, which we will discuss in later chapters, or by *rejection sampling* (see Challenge 2.10.21).

Summary of Section 2.10

- It is important to be able to simulate probability distributions.
- If X is discrete, taking the value x_i with probability p_i, where $x_1 < x_2 < \cdots$, and $U \sim$ Uniform[0, 1], and $Y = \min\{x_j : \sum_{k=1}^{j} p_k \geq U\}$, then Y has the same distribution as X. This method can be used to simulate virtually any discrete distribution.
- If F is any cumulative distribution with inverse cdf F^{-1}, $U \sim$ Uniform[0, 1], and $Y = F^{-1}(U)$, then Y has cumulative distribution function F. This allows us to simulate virtually any continuous distribution.
- There are simple methods of simulating many standard distributions, including the binomial, uniform, exponential, and normal.

EXERCISES

2.10.1 Let Y be a discrete random variable with $P(Y = -7) = 1/2$, $P(Y = -2) = 1/3$, and $P(Y = 5) = 1/6$. Find a formula for Z in terms of U, such that if $U \sim$ Uniform[0, 1], then Z has the same distribution as Y.

2.10.2 For each of the following cumulative distribution functions F, find a formula for X in terms of U, such that if $U \sim$ Uniform[0, 1], then X has cumulative distribution function F.

(a)

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

(b)

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2 & 0 \le x \le 1 \\ 1 & x > 1 \end{cases}$$

(c)

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2/9 & 0 \le x \le 3 \\ 1 & x > 3 \end{cases}$$

(d)

$$F(x) = \begin{cases} 0 & x < 1 \\ x^2/9 & 1 \le x \le 3 \\ 1 & x > 3 \end{cases}$$

(e)

$$F(x) = \begin{cases} 0 & x < 0 \\ x^5/32 & 0 \le x \le 2 \\ 1 & x > 2 \end{cases}$$

(f)

$$F(x) = \begin{cases} 0 & x < 0 \\ 1/3 & 0 \le x < 7 \\ 3/4 & 7 \le x < 11 \\ 1 & x \ge 11 \end{cases}$$

2.10.3 Suppose $U \sim$ Uniform[0, 1], and $Y = \ln(1/U) \,/\, 3$. What is the distribution of Y?

2.10.4 Generalizing the previous question, suppose $U \sim$ Uniform[0, 1] and $W = \ln(1/U) \,/\, \lambda$ for some fixed $\lambda > 0$.

(a) What is the distribution of W?

(b) Does this provide a way of simulating from a certain well-known distribution? Explain.

2.10.5 Let $U_1 \sim$ Uniform$[0, 1]$ and $U_2 \sim$ Uniform$[0, 1]$ be independent, and let $X = c_1\sqrt{\log(1/U_1)}\,\cos(2\pi U_2) + c_2$. Find values of c_1 and c_2 such that $X \sim N(5, 9)$.

2.10.6 Let $U \sim$ Uniform$[0, 1]$. Find a formula for Y in terms of U, such that $P(Y = 3) = P(Y = 4) = 2/5$ and $P(Y = 7) = 1/5$, otherwise $P(Y = y) = 0$.

2.10.7 Suppose $P(X = 1) = 1/3$, $P(X = 2) = 1/6$, $P(X = 4) = 1/2$, and $P(X = x) = 0$ otherwise.
(a) Compute the cdf $F_X(x)$ for all $x \in R^1$.
(b) Compute the inverse cdf $F_X^{-1}(t)$ for all $t \in R^1$.
(c) Let $U \sim$ Uniform$[0, 1]$. Find a formula for Y in terms of U, such that Y has cdf F_X.

2.10.8 Let X have density function $f_X(x) = 3\sqrt{x}/2$ for $0 < x < 1$, otherwise $f_X(x) = 0$.
(a) Compute the cdf $F_X(x)$ for all $x \in R^1$.
(b) Compute the inverse cdf $F_X^{-1}(t)$ for all $t \in R^1$.
(c) Let $U \sim$ Uniform$[0, 1]$. Find a formula for Y in terms of U, such that Y has density f.

2.10.9 Let $U \sim$ Uniform$[0, 1]$. Find a formula for Z in terms of U, such that Z has density $f_Z(z) = 4z^3$ for $0 < z < 1$, otherwise $f_Z(z) = 0$.

COMPUTER EXERCISES

2.10.10 For each of the following distributions, use the computer (you can use any algorithms available to you as part of a software package) to simulate $X_1, X_2, \ldots, X_N$ i.i.d. having the given distribution. (Take $N = 1000$ at least, with $N = 10{,}000$ or $N = 100{,}000$ if possible.) Then compute $\bar{X} = (1/N)\sum_{i=1}^{N} X_i$ and $(1/N)\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$.
(a) Uniform$[0, 1]$
(b) Uniform$[5, 8]$
(c) Bernoulli$(1/3)$
(d) Binomial$(12,\ 1/3)$
(e) Geometric$(1/5)$
(f) Exponential(1)
(g) Exponential(13)
(h) $N(0, 1)$
(i) $N(5, 9)$

PROBLEMS

2.10.11 Let $G(x) = p_1F_1(x) + p_2F_2(x) + \cdots + p_kF_k(x)$, where $p_i \geq 0$, $\sum_i p_i = 1$, and F_i are cdfs, as in (2.5.3). Suppose we can generate X_i to have cdf F_i, for $i = 1, 2, \ldots, k$. Describe a procedure for generating a random variable Y that has cdf G.

2.10.12 Let X be an absolutely continuous random variable, with density given by $f_X(x) = x^{-2}$ for $x \geq 1$, with $f_X(x) = 0$ otherwise. Find a formula for Z in terms of U, such that if $U \sim$ Uniform$[0, 1]$, then Z has the same distribution as X.

2.10.13 Find the inverse cdf of the logistic distribution of Problem 2.4.18. (Hint: See Problem 2.5.20.)

2.10.14 Find the inverse cdf of the Weibull(α) distribution of Problem 2.4.19. (Hint: See Problem 2.5.21.)

2.10.15 Find the inverse cdf of the Pareto(α) distribution of Problem 2.4.20. (Hint: See Problem 2.5.22.)

2.10.16 Find the inverse cdf of the Cauchy distribution of Problem 2.4.21. (Hint: See Problem 2.5.23.)

2.10.17 Find the inverse cdf of the Laplace distribution of Problem 2.4.22. (Hint: See Problem 2.5.24.)

2.10.18 Find the inverse cdf of the extreme value distribution of Problem 2.4.23. (Hint: See Problem 2.5.25.)

2.10.19 Find the inverse cdfs of the beta distributions in Problem 2.4.24(b) through (d). (Hint: See Problem 2.5.26.)

2.10.20 (*Method of composition*) If we generate $X \sim f_X$ obtaining x, and then generate Y from $f_{Y|X}(\cdot \,|\, x)$, prove that $Y \sim f_Y$.

CHALLENGES

2.10.21 (*Rejection sampling*) Suppose f is a complicated density function. Suppose g is a density function from which it is easy to sample (e.g., the density of a uniform or exponential or normal distribution). Suppose we know a value of c such that $f(x) \le cg(x)$ for all $x \in R^1$. The following provides a method, called rejection sampling, for sampling from a complicated density f by using a simpler density g, provided only that we know $f(x) \le cg(x)$ for all $x \in R^1$.

(a) Suppose Y has density g. Let $U \sim$ Uniform$[0, c]$, with U and Y independent. Prove that

$$P(a \le Y \le b \,|\, f(Y) \ge Ucg(Y)) = \int_a^b f(x)\,dx.$$

(Hint: Use Theorem 2.8.1 to show that $P(a \le Y \le b,\ f(Y) \ge cUg(Y)) = \int_a^b g(y)P(f(Y) \ge cUg(Y) \,|\, Y = y)\,dy$.)

(b) Suppose that $Y_1, Y_2, \ldots$ are i.i.d., each with density g, and independently $U_1, U_2, \ldots$ are i.i.d. Uniform$[0, c]$. Let $i_0 = 0$, and for $n \ge 1$, let $i_n = \min\{j > i_{n-1} : U_j f(Y_j) \ge cg(Y_j)\}$. Prove that $X_{i_1}, X_{i_2}, \ldots$ are i.i.d., each with density f. (Hint: Prove this for X_{i_1}, X_{i_2}.)

2.11 Further Proofs (Advanced)

Proof of Theorem 2.4.2

We want to prove that the function ϕ given by (2.4.9) is a density function.

Clearly $\phi(x) \geq 0$ for all x. To proceed, we set $I = \int_{-\infty}^{\infty} \phi(x)\,dx$. Then, using multivariable calculus,

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} \phi(x)\,dx\right)^2 = \left(\int_{-\infty}^{\infty} \phi(x)\,dx\right)\left(\int_{-\infty}^{\infty} \phi(y)\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \phi(x)\,\phi(y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}\,e^{-(x^2+y^2)/2}\,dx\,dy. \end{aligned}$$

We now switch to polar coordinates (r, θ), so that $x = r\cos\theta$ and $y = r\sin\theta$, where $r > 0$ and $0 \leq \theta \leq 2\pi$. Then $x^2 + y^2 = r^2$ and, by the multivariable change of variable theorem from calculus, $dx\,dy = r\,dr\,d\theta$. Hence,

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} \frac{1}{2\pi} e^{-r^2/2} r\,dr\,d\theta = \int_0^{\infty} e^{-r^2/2} r\,dr \\ &= -e^{-r^2/2}\Big|_{r=0}^{r=\infty} = (-0) - (-1) = 1, \end{aligned}$$

and we have $I^2 = 1$. But clearly $I \geq 0$ (because $\phi \geq 0$), so we must have $I = 1$, as claimed. ▮

Proof of Theorem 2.6.2

We want to prove that, when X is an absolutely continuous random variable, with density function f_X and $Y = h(X)$, where $h : R^1 \to R^1$ is a function that is differentiable and strictly increasing, then Y is also absolutely continuous, and its density function f_Y is given by

$$f_Y(y) = f_X(h^{-1}(y)) \,/\, |h'(h^{-1}(y))|, \tag{2.11.1}$$

where h' is the derivative of h, and where $h^{-1}(y)$ is the unique number x such that $h(x) = y$.

We must show that whenever $a \leq b$, we have

$$P(a \leq Y \leq b) = \int_a^b f_Y(y)\,dy,$$

where f_Y is given by (2.11.1). To that end, we note that, because h is strictly increasing, so is h^{-1}. Hence, applying h^{-1} preserves inequalities, so that

$$\begin{aligned} P(a \leq Y \leq b) &= P(h^{-1}(a) \leq h^{-1}(Y) \leq h^{-1}(b)) = P(h^{-1}(a) \leq X \leq h^{-1}(b)) \\ &= \int_{h^{-1}(a)}^{h^{-1}(b)} f_X(x)\,dx. \end{aligned}$$

We then make the substitution $y = h(x)$, so that $x = h^{-1}(y)$, and

$$dx = \left|\frac{d}{dy}h^{-1}(y)\right| dy.$$

But by the inverse function theorem from calculus, $\frac{d}{dy}h^{-1}(y) = 1/h'(h^{-1}(y))$. Furthermore, as x goes from $h^{-1}(a)$ to $h^{-1}(b)$, we see that $y = h(x)$ goes from a to b. We conclude that

$$\begin{aligned} P(a \le Y \le b) &= \int_{h^{-1}(a)}^{h^{-1}(b)} f_X(x)\,dx = \int_a^b f_X(h^{-1}(y))(1/|h'(h^{-1}(y))|)\,dy \\ &= \int_a^b f_Y(y)\,dy, \end{aligned}$$

as required. ▮

Proof of Theorem 2.6.3

We want to prove that when X is an absolutely continuous random variable, with density function f_X and $Y = h(X)$, where $h : R^1 \to R^1$ is a function that is differentiable and strictly decreasing, then Y is also absolutely continuous, and its density function f_Y may again be defined by (2.11.1).

We note that, because h is strictly decreasing, so is h^{-1}. Hence, applying h^{-1} reverses the inequalities, so that

$$\begin{aligned} P(a \le Y \le b) &= P(h^{-1}(b) \le h^{-1}(Y) \le h^{-1}(a)) = P(h^{-1}(b) \le X \le h^{-1}(a)) \\ &= \int_{h^{-1}(b)}^{h^{-1}(a)} f_X(x)\,dx. \end{aligned}$$

We then make the substitution $y = h(x)$, so that $x = h^{-1}(y)$, and

$$dx = \left|\frac{d}{dy}h^{-1}(y)\right| dy.$$

But by the inverse function theorem from calculus,

$$\frac{d}{dy}h^{-1}(y) = \frac{1}{h'(h^{-1}(y))}.$$

Furthermore, as x goes from $h^{-1}(b)$ to $h^{-1}(a)$, we see that $y = h(x)$ goes from a to b. We conclude that

$$\begin{aligned} P(a \le Y \le b) &= \int_{h^{-1}(b)}^{h^{-1}(a)} f_X(x)\,dx = \int_a^b f_X(h^{-1}(y))\,(1/|h'(h^{-1}(y))|)\,dy \\ &= \int_a^b f_Y(y)\,dy, \end{aligned}$$

as required. ▮

Proof of Theorem 2.9.2

We want to prove the following result. Let X and Y be jointly absolutely continuous, with joint density function $f_{X,Y}$. Let $Z = h_1(X, Y)$ and $W = h_2(X, Y)$, where $h_1, h_2 : R^2 \to R^1$ are differentiable functions. Define the joint function $h = (h_1, h_2) : R^2 \to R^2$ by

$$h(x, y) = (h_1(x, y), h_2(x, y))\,.$$

Assume that h is one-to-one, at least on the region $\{(x, y) : f(x, y) > 0\}$, i.e., if $h_1(x_1, y_1) = h_1(x_2, y_2)$ and $h_2(x_1, y_1) = h_2(x_2, y_2)$, then $x_1 = x_2$ and $y_1 = y_2$. Then Z and W are also jointly absolutely continuous, with joint density function $f_{Z,W}$ given by

$$f_{Z,W}(z, w) = f_{X,Y}(h^{-1}(z, w)) \,/\, |J(h^{-1}(z, w))|,$$

where J is the Jacobian derivative of h, and where $h^{-1}(z, w)$ is the unique pair (x, y) such that $h(x, y) = (z, w)$.

We must show that whenever $a \le b$ and $c \le d$, we have

$$P(a \le Z \le b\,,\ c \le W \le d) = \int_c^d \int_a^b f_{Z,W}(z, w)\, dw\, dz.$$

If we let $S = [a, b] \times [c, d]$ be the two-dimensional rectangle, then we can rewrite this as

$$P((Z, W) \in S) = \int\int_S f_{Z,W}(z, w)\, dz\, dw.$$

Now, using the theory of multivariable calculus, and making the substitution $(x, y) = h^{-1}(z, w)$ (which is permissible because h is one-to-one), we have

$$\begin{aligned}
&\int\int_S f_{Z,W}(z, w)\, dz\, dw \\
&= \int\int_S \Big(f_{X,Y}(h^{-1}(z, w)) \,/\, |J(h^{-1}(z, w))|\Big)\, dz\, dw \\
&= \int\int_{h^{-1}(S)} \big(f_{X,Y}(x, y) \,/\, |J(x, y)|\big)\, |J(x, y)|\, dx\, dy \\
&= \int\int_{h^{-1}(S)} f_{X,Y}(x, y)\, dx\, dy = P((X, Y) \in h^{-1}(S)) \\
&= P(h^{-1}(Z, W) \in h^{-1}(S)) \ = P((Z, W) \in S),
\end{aligned}$$

as required. ∎