Chapter 4

Multiple Random Variables

"I confess that I have been blind as a mole, but it is better to learn wisdom late than never to learn it at all."

Sherlock Holmes
The Man with the Twisted Lip

4.1 Joint and Marginal Distributions

In previous chapters, we have discussed probability models and computation of probability for events involving only one random variable. These are called *univariate models.* In this chapter, we discuss probability models that involve more than one random variable—naturally enough, called *multivariate models.*

In an experimental situation, it would be very unusual to observe only the value of one random variable. That is, it would be an unusual experiment in which the total data collected consisted of just one numeric value. For example, consider an experiment designed to gain information about some health characteristics of a population of people. It would be a modest experiment indeed if the only datum collected was the body weight of one person. Rather, the body weights of several people in the population might be measured. These different weights would be observations on different random variables, one for each person measured. Multiple observations could also arise because several physical characteristics were measured on each person. For example, temperature, height, and blood pressure, in addition to weight, might be measured. These observations on different characteristics could also be modeled as observations on different random variables. Thus, we need to know how to describe and use probability models that deal with more than one random variable at a time. For the first several sections we will discuss mainly *bivariate models,* models involving two random variables.

Recall that, in Definition 1.4.1, a (univariate) random variable was defined to be a function from a sample space S into the real numbers. A random vector, consisting of several random variables, is defined similarly.

Definition 4.1.1 An *n-dimensional random vector* is a function from a sample space S into $\Re^n$, n-dimensional Euclidean space.

Suppose, for example, that with each point in a sample space we associate an ordered pair of numbers, that is, a point $(x, y) \in \Re^2$, where $\Re^2$ denotes the plane. Then

we have defined a two-dimensional (or bivariate) random vector (X, Y). Example 4.1.2 illustrates this.

Example 4.1.2 (Sample space for dice) Consider the experiment of tossing two fair dice. The sample space for this experiment has 36 equally likely points and was introduced in Example 1.3.10. For example, the sample point $(3, 3)$ denotes the outcome in which both dice show a 3; the sample point $(4, 1)$ denotes the outcome in which the first die shows a 4 and the second die a 1; etc. Now, with each of these 36 points associate two numbers, X and Y. Let

$$X = \text{ sum of the two dice} \quad \text{and} \quad Y = |\text{difference of the two dice}|.$$

For the sample point $(3, 3)$, $X = 3+3 = 6$ and $Y = |3 - 3| = 0$. For $(4, 1)$, $X = 5$ and $Y = 3$. These are also the values of X and Y for the sample point $(1, 4)$. For each of the 36 sample points we could compute the values of X and Y. In this way we have defined the bivariate random vector (X, Y).

Having defined a random vector (X, Y), we can now discuss probabilities of events that are defined in terms of (X, Y). The probabilities of events defined in terms of X and Y are just defined in terms of the probabilities of the corresponding events in the sample space S. What is $P(X = 5 \text{ and } Y = 3)$? You can verify that the only two sample points that yield $X = 5$ and $Y = 3$ are $(4, 1)$ and $(1, 4)$. Thus the event "$X = 5$ and $Y = 3$" will occur if and only if the event $\{(4, 1), (1, 4)\}$ occurs. Since each of the 36 sample points in S is equally likely,

$$P(\{(4, 1), (1, 4)\}) = \frac{2}{36} = \frac{1}{18}.$$

Thus,

$$P(X = 5 \text{ and } Y = 3) = \frac{1}{18}.$$

Henceforth, we will write $P(X = 5, Y = 3)$ for $P(X = 5 \text{ and } Y = 3)$. Read the comma as "and." Similarly, $P(X = 6, Y = 0) = \frac{1}{36}$ because the only sample point that yields these values of X and Y is $(3, 3)$. For more complicated events, the technique is the same. For example, $P(X = 7, Y \leq 4) = \frac{4}{36} = \frac{1}{9}$ because the only four sample points that yield $X = 7$ and $Y \leq 4$ are $(4, 3)$, $(3, 4)$, $(5, 2)$, and $(2, 5)$. ‖

The random vector (X, Y) defined above is called a *discrete random vector* because it has only a countable (in this case, finite) number of possible values. For a discrete random vector, the function $f(x, y)$ defined by $f(x, y) = P(X = x, Y = y)$ can be used to compute any probabilities of events defined in terms of (X, Y).

Definition 4.1.3 Let (X, Y) be a discrete bivariate random vector. Then the function $f(x, y)$ from $\Re^2$ into $\Re$ defined by $f(x, y) = P(X = x, Y = y)$ is called the *joint probability mass function* or *joint pmf* of (X, Y). If it is necessary to stress the fact that f is the joint pmf of the vector (X, Y) rather than some other vector, the notation $f_{X,Y}(x, y)$ will be used.

		x									
		2	3	4	5	6	7	8	9	10	11	12
	0	$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$
	1		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$	
y	2			$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		
	3				$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$			
	4					$\frac{1}{18}$		$\frac{1}{18}$				
	5						$\frac{1}{18}$					

Table 4.1.1. *Values of the joint pmf $f(x,y)$*

The joint pmf of (X,Y) completely defines the probability distribution of the random vector (X,Y), just as the pmf of a discrete univariate random variable completely defines its distribution. For the (X,Y) defined in Example 4.1.2 in terms of the roll of a pair of dice, there are 21 possible values of (X,Y). The value of $f(x,y)$ for each of these 21 possible values is given in Table 4.1.1. Two of these values, $f(5,3) = \frac{1}{18}$ and $f(6,0) = \frac{1}{36}$, were computed above and the rest are obtained by similar reasoning. The joint pmf $f(x,y)$ is defined for all $(x,y) \in \Re^2$, not just the 21 pairs in Table 4.1.1. For any other (x,y), $f(x,y) = P(X = x, Y = y) = 0$.

The joint pmf can be used to compute the probability of any event defined in terms of (X,Y). Let A be any subset of $\Re^2$. Then

$$P((X,Y) \in A) = \sum_{(x,y)\in A} f(x,y).$$

Since (X,Y) is discrete, $f(x,y)$ is nonzero for at most a countable number of points (x,y). Thus, the sum can be interpreted as a countable sum even if A contains an uncountable number of points. For example, let $A = \{(x,y) : x = 7 \text{ and } y \le 4\}$. This is a half-infinite line in $\Re^2$. But from Table 4.1.1 we see that the only $(x,y) \in A$ for which $f(x,y)$ is nonzero are $(x,y) = (7,1)$ and $(x,y) = (7,3)$. Thus,

$$P(X = 7, Y \le 4) = P((X,Y) \in A) = f(7,1) + f(7,3) = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

This, of course, is the same value computed in Example 4.1.2 by considering the definition of (X,Y) and sample points in S. It is usually simpler to work with the joint pmf than it is to work with the fundamental definition.

Expectations of functions of random vectors are computed just as with univariate random variables. Let $g(x,y)$ be a real-valued function defined for all possible values (x,y) of the discrete random vector (X,Y). Then $g(X,Y)$ is itself a random variable and its expected value $Eg(X,Y)$ is given by

$$Eg(X,Y) = \sum_{(x,y)\in\Re^2} g(x,y)f(x,y).$$

Example 4.1.4 (Continuation of Example 4.1.2) For the (X,Y) whose joint pmf is given in Table 4.1.1, what is the average value of XY? Letting $g(x,y) = xy$, we

compute $EXY = Eg(X,Y)$ by computing $xyf(x,y)$ for each of the 21 (x,y) points in Table 4.1.1 and summing these 21 terms. Thus,

$$EXY = (2)(0)\frac{1}{36} + (4)(0)\frac{1}{36} + \cdots + (8)(4)\frac{1}{18} + (7)(5)\frac{1}{18} = 13\frac{11}{18}. \quad \|$$

The expectation operator continues to have the properties listed in Theorem 2.2.5 when the random variable X is replaced by the random vector (X,Y). For example, if $g_1(x,y)$ and $g_2(x,y)$ are two functions and a, b, and c are constants, then

$$E(ag_1(X,Y) + bg_2(X,Y) + c) = aEg_1(X,Y) + bEg_2(X,Y) + c.$$

These properties follow from the properties of sums exactly as in the univariate case (see Exercise 4.2).

The joint pmf for any discrete bivariate random vector (X,Y) must have certain properties. For any (x,y), $f(x,y) \geq 0$ since $f(x,y)$ is a probability. Also, since (X,Y) is certain to be in $\Re^2$,

$$\sum_{(x,y)\in\Re^2} f(x,y) = P\left((X,Y) \in \Re^2\right) = 1.$$

It turns out that any nonnegative function from $\Re^2$ into $\Re$ that is nonzero for at most a countable number of (x,y) pairs and sums to 1 is the joint pmf for some bivariate discrete random vector (X,Y). Thus, by defining $f(x,y)$, we can define a probability model for (X,Y) without ever working with the fundamental sample space S.

Example 4.1.5 (Joint pmf for dice) Define $f(x,y)$ by

$$f(0,0) = f(0,1) = \frac{1}{6},$$
$$f(1,0) = f(1,1) = \frac{1}{3},$$
$$f(x,y) = 0 \quad \text{for any other } (x,y).$$

Then $f(x,y)$ is nonnegative and sums to 1, so $f(x,y)$ is the joint pmf for some bivariate random vector (X,Y). We can use $f(x,y)$ to compute probabilities such as $P(X = Y) = f(0,0) + f(1,1) = \frac{1}{2}$. All this can be done without reference to the sample space S. Indeed, there are many sample spaces and functions thereon that lead to this joint pmf for (X,Y). Here is one. Let S be the 36-point sample space for the experiment of tossing two fair dice. Let $X = 0$ if the first die shows at most 2 and $X = 1$ if the first die shows more than 2. Let $Y = 0$ if the second die shows an odd number and $Y = 1$ if the second die shows an even number. It is left as Exercise 4.3 to show that this definition leads to the above probability distribution for (X,Y). $\|$

Even if we are considering a probability model for a random vector (X,Y), there may be probabilities or expectations of interest that involve only one of the random variables in the vector. We may wish to know $P(X = 2)$, for instance. The variable X is itself a random variable, in the sense of Chapter 1, and its probability distribution

is described by its pmf, namely, $f_X(x) = P(X = x)$. (As mentioned earlier, we now use the subscript to distinguish $f_X(x)$ from the joint pmf $f_{X,Y}(x,y)$.) We now call $f_X(x)$ the *marginal pmf of* X to emphasize the fact that it is the pmf of X but in the context of the probability model that gives the joint distribution of the vector (X,Y). The marginal pmf of X or Y is easily calculated from the joint pmf of (X,Y) as Theorem 4.1.6 indicates.

Theorem 4.1.6 *Let (X,Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x,y)$. Then the marginal pmfs of X and Y, $f_X(x) = P(X = x)$ and $f_Y(y) = P(Y = y)$, are given by*

$$f_X(x) = \sum_{y \in \Re} f_{X,Y}(x,y) \quad \text{and} \quad f_Y(y) = \sum_{x \in \Re} f_{X,Y}(x,y).$$

Proof: We will prove the result for $f_X(x)$. The proof for $f_Y(y)$ is similar. For any $x \in \Re$, let $A_x = \{(x,y) : -\infty < y < \infty\}$. That is, A_x is the line in the plane with first coordinate equal to x. Then, for any $x \in \Re$,

$$\begin{aligned} f_X(x) &= P(X = x) \\ &= P(X = x, -\infty < Y < \infty) && (P(-\infty < Y < \infty) = 1) \\ &= P((X,Y) \in A_x) && (\text{definition of } A_x) \\ &= \sum_{(x,y) \in A_x} f_{X,Y}(x,y) \\ &= \sum_{y \in \Re} f_{X,Y}(x,y). \end{aligned}$$

□

Example 4.1.7 (Marginal pmf for dice) Using the result of Theorem 4.1.6, we can compute the marginal distributions for X and Y from the joint distribution given in Table 4.1.1. To compute the marginal pmf of Y, for each possible value of Y we sum over the possible values of X. In this way we obtain

$$\begin{aligned} f_Y(0) &= f_{X,Y}(2,0) + f_{X,Y}(4,0) + f_{X,Y}(6,0) \\ &\quad + f_{X,Y}(8,0) + f_{X,Y}(10,0) + f_{X,Y}(12,0) \\ &= \frac{1}{6}. \end{aligned}$$

Similarly, we obtain

$$f_Y(1) = \tfrac{5}{18}, \quad f_Y(2) = \tfrac{2}{9}, \quad f_Y(3) = \tfrac{1}{6}, \quad f_Y(4) = \tfrac{1}{9}, \quad f_Y(5) = \tfrac{1}{18}.$$

Notice that $f_Y(0) + f_Y(1) + f_Y(2) + f_Y(3) + f_Y(4) + f_Y(5) = 1$, as it must, since these are the only six possible values of Y. ||

The marginal pmf of X or Y is the same as the pmf of X or Y defined in Chapter 1. The marginal pmf of X or Y can be used to compute probabilities or expectations that involve only X or Y. But to compute a probability or expectation that simultaneously involves both X and Y, we must use the joint pmf of X and Y.

Example 4.1.8 (Dice probabilities) Using the marginal pmf of Y computed in Example 4.1.7, we can compute

$$P(Y < 3) = f_Y(0) + f_Y(1) + f_Y(2) = \frac{1}{6} + \frac{5}{18} + \frac{2}{9} = \frac{2}{3}.$$

Also,

$$\mathrm{E}Y^3 = 0^3 f_Y(0) + \cdots + 5^3 f_Y(5) = 20\frac{11}{18}. \quad \|$$

The marginal distributions of X and Y, described by the marginal pmfs $f_X(x)$ and $f_Y(y)$, do not completely describe the joint distribution of X and Y. Indeed, there are many different joint distributions that have the same marginal distributions. Thus, it is hopeless to try to determine the joint pmf, $f_{X,Y}(x,y)$, from knowledge of only the marginal pmfs, $f_X(x)$ and $f_Y(y)$. The next example illustrates the point.

Example 4.1.9 (Same marginals, different joint pmf) Define a joint pmf by

$$f(0,0) = \tfrac{1}{12}, \quad f(1,0) = \tfrac{5}{12}, \quad f(0,1) = f(1,1) = \tfrac{3}{12},$$
$$f(x,y) = 0 \quad \text{for all other values.}$$

The marginal pmf of Y is $f_Y(0) = f(0,0) + f(1,0) = \frac{1}{2}$ and $f_Y(1) = f(0,1) + f(1,1) = \frac{1}{2}$. The marginal pmf of X is $f_X(0) = \frac{1}{3}$ and $f_X(1) = \frac{2}{3}$. Now check that for the joint pmf given in Example 4.1.5, which is obviously different from the one given here, the marginal pmfs of both X and Y are exactly the same as the ones just computed. Thus, we cannot determine what the joint pmf is if we know only the marginal pmfs. The joint pmf tells us additional information about the distribution of (X,Y) that is not found in the marginal distributions. $\|$

To this point we have discussed discrete bivariate random vectors. We can also consider random vectors whose components are continuous random variables. The probability distribution of a continuous random vector is usually described using a density function, as in the univariate case.

Definition 4.1.10 A function $f(x,y)$ from $\Re^2$ into $\Re$ is called a *joint probability density function* or *joint pdf of the continuous bivariate random vector* (X,Y) if, for every $A \subset \Re^2$,

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

A joint pdf is used just like a univariate pdf except now the integrals are double integrals over sets in the plane. The notation $\int\int_A$ simply means that the limits of integration are set so that the function is integrated over all $(x,y) \in A$. Expectations of functions of continuous random vectors are defined as in the discrete case with integrals replacing sums and the pdf replacing the pmf. That is, if $g(x,y)$ is a real-valued function, then the *expected value of* $g(X,Y)$ is defined to be

$$\mathrm{E}g(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\,dx\,dy. \tag{4.1.2}$$

It is important to realize that the joint pdf *is* defined for all $(x, y) \in \Re^2$. The pdf may equal 0 on a large set A if $P((X, Y) \in A) = 0$ but the pdf *is* defined for the points in A.

The *marginal probability density functions* of X and Y are also defined as in the discrete case with integrals replacing sums. The marginal pdfs may be used to compute probabilities or expectations that involve only X or Y. Specifically, the marginal pdfs of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad -\infty < x < \infty,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx, \quad -\infty < y < \infty. \tag{4.1.3}$$

Any function $f(x, y)$ satisfying $f(x, y) \geq 0$ for all $(x, y) \in \Re^2$ and

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

is the joint pdf of some continuous bivariate random vector (X, Y). All of these concepts regarding joint pdfs are illustrated in the following two examples.

Example 4.1.11 (Calculating joint probabilities–I) Define a joint pdf by

$$f(x, y) = \begin{cases} 6xy^2 & 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(Henceforth, it will be understood that $f(x, y) = 0$ for (x, y) values not specifically mentioned in the definition.) First, we might check that $f(x, y)$ is indeed a joint pdf. That $f(x, y) \geq 0$ for all (x, y) in the defined range is fairly obvious. To compute the integral of $f(x, y)$ over the whole plane, note that, since $f(x, y)$ is 0 except on the unit square, the integral over the plane is the same as the integral over the square. Thus we have

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = \int_0^1 \int_0^1 6xy^2\, dx\, dy = \int_0^1 3x^2y^2 \Big|_0^1 dy$$
$$= \int_0^1 3y^2\, dy = y^3 \Big|_0^1 = 1.$$

Now, consider calculating a probability such as $P(X + Y \geq 1)$. Letting $A = \{(x, y) : x + y \geq 1\}$, we can re-express this as $P((X, Y) \in A)$. From Definition 4.1.10, to calculate the probability we integrate the joint pdf over the set A. But the joint pdf is 0 except on the unit square. So integrating over A is the same as integrating over only that part of A which is in the unit square. The set A is a half-plane in the northeast part of the plane, and the part of A in the unit square is the triangular region bounded by the lines $x = 1$, $y = 1$, and $x + y = 1$. We can write

$$\begin{aligned} A &= \{(x,y) : x+y \geq 1, 0 < x < 1, 0 < y < 1\} \\ &= \{(x,y) : x \geq 1-y, 0 < x < 1, 0 < y < 1\} \\ &= \{(x,y) : 1 - y \leq x < 1, 0 < y < 1\}. \end{aligned}$$

This gives us the limits of integration we need to calculate the probability. We have

$$P(X+Y \geq 1) = \int_A \int f(x,y)\,dx\,dy = \int_0^1 \int_{1-y}^1 6xy^2\,dx\,dy = \frac{9}{10}.$$

Using (4.1.3), we can calculate the marginal pdf of X or Y. For example, to calculate $f_X(x)$, we note that for $x \geq 1$ or $x \leq 0$, $f(x,y) = 0$ for all values of y. Thus for $x \geq 1$ or $x \leq 0$,

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

For $0 < x < 1$, $f(x,y)$ is nonzero only if $0 < y < 1$. Thus for $0 < x < 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 6xy^2\,dy = 2xy^3\Big|_0^1 = 2x.$$

This marginal pdf of X can now be used to calculate probabilities involving only X. For example,

$$P\left(\frac{1}{2} < X < \frac{3}{4}\right) = \int_{\frac{1}{2}}^{\frac{3}{4}} 2x\,dx = \frac{5}{16}. \qquad \|$$

Example 4.1.12 (Calculating joint probabilities–II) As another example of a joint pdf, let $f(x,y) = e^{-y}, 0 < x < y < \infty$. Although e^{-y} does not depend on x, $f(x,y)$ certainly is a function of x since the set where $f(x,y)$ is nonzero depends on x. This is made more obvious by using an indicator function to write

$$f(x,y) = e^{-y} I_{\{(u,v):0<u<v<\infty\}}(x,y).$$

To calculate $P(X+Y \geq 1)$, we could integrate the joint pdf over the region that is the intersection of the set $A = \{(x,y) : x+y \geq 1\}$ and the set where $f(x,y)$ is nonzero. Graph these sets and notice that this region is an unbounded region (lighter shading in Figure 4.1.1) with three sides given by the lines $x = y, x+y = 1$, and $x = 0$. To integrate over this region we would have to break the region into at least two parts in order to write the appropriate limits of integration.

The integration is easier over the intersection of the set $B = \{(x,y) : x+y < 1\}$ and the set where $f(x,y)$ is nonzero, the triangular region (darker shading in Figure 4.1.1) bounded by the lines $x = y$, $x+y = 1$, and $x = 0$. Thus

$$\begin{aligned} P(X+Y \geq 1) &= 1 - P(X+Y < 1) = 1 - \int_0^{\frac{1}{2}} \int_x^{1-x} e^{-y}\,dy\,dx \\ &= 1 - \int_0^{\frac{1}{2}} (e^{-x} - e^{-(1-x)})\,dx = 2e^{-1/2} - e^{-1}. \end{aligned}$$

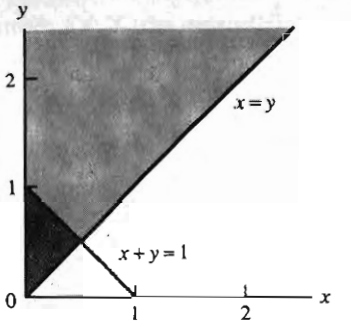

Figure 4.1.1. *Regions for Example 4.1.12*

This illustrates that it is almost always helpful to graph the sets of interest in determining the appropriate limits of integration for problems such as this. ||

The joint probability distribution of (X, Y) can be completely described with the *joint cdf* (cumulative distribution function) rather than with the joint pmf or joint pdf. The joint cdf is the function $F(x, y)$ defined by

$$F(x, y) = P(X \leq x, Y \leq y)$$

for all $(x, y) \in \Re^2$. The joint cdf is usually not very handy to use for a discrete random vector. But for a continuous bivariate random vector we have the important relationship, as in the univariate case,

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, dt\, ds.$$

From the bivariate Fundamental Theorem of Calculus, this implies that

$$\frac{\partial^2 F(x, y)}{\partial x\, \partial y} = f(x, y) \tag{4.1.4}$$

at continuity points of $f(x, y)$. This relationship is useful in situations where an expression for $F(x, y)$ can be found. The mixed partial derivative can be computed to find the joint pdf.

4.2 Conditional Distributions and Independence

Oftentimes when two random variables, (X, Y), are observed, the values of the two variables are related. For example, suppose that, in sampling from a human population, X denotes a person's height and Y denotes the same person's weight. Surely we would think it more likely that $Y > 200$ pounds if we were told that $X = 73$ inches than if we were told that $X = 41$ inches. Knowledge about the value of X gives us some information about the value of Y even if it does not tell us the value of Y exactly. Conditional probabilities regarding Y given knowledge of the X value can

be computed using the joint distribution of (X, Y). Sometimes, however, knowledge about X gives us no information about Y. We will discuss these topics concerning conditional probabilities in this section.

If (X, Y) is a discrete random vector, then a conditional probability of the form $P(Y = y|X = x)$ is interpreted exactly as in Definition 1.3.2. For a countable (maybe finite) number of x values, $P(X = x) > 0$. For these values of x, $P(Y = y|X = x)$ is simply $P(X = x, Y = y)/P(X = x)$, according to the definition. The event $\{Y = y\}$ is the event A in the formula and the event $\{X = x\}$ is the event B. For a fixed value of x, $P(Y = y|X = x)$ could be computed for all possible values of y. In this way the probability of various values of y could be assessed given the knowledge that $X = x$ was observed. This computation can be simplified by noting that in terms of the joint and marginal pmfs of X and Y, the above probabilities are $P(X = x, Y = y) = f(x, y)$ and $P(X = x) = f_X(x)$. This leads to the following definition.

Definition 4.2.1 Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X = x) = f_X(x) > 0$, the *conditional pmf of* Y *given that* $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = P(Y = y|X = x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $P(Y = y) = f_Y(y) > 0$, the *conditional pmf of* X *given that* $Y = y$ is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = P(X = x|Y = y) = \frac{f(x, y)}{f_Y(y)}.$$

Since we have called $f(y|x)$ a pmf, we should verify that this function of y does indeed define a pmf for a random variable. First, $f(y|x) \geq 0$ for every y since $f(x, y) \geq 0$ and $f_X(x) > 0$. Second,

$$\sum_y f(y|x) = \frac{\sum_y f(x, y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$$

Thus, $f(y|x)$ is indeed a pmf and can be used in the usual way to compute probabilities involving Y given the knowledge that $X = x$ occurred.

Example 4.2.2 (Calculating conditional probabilities) Define the joint pmf of (X, Y) by

$$f(0, 10) = f(0, 20) = \tfrac{2}{18}, \quad f(1, 10) = f(1, 30) = \tfrac{3}{18},$$
$$f(1, 20) = \tfrac{4}{18}, \quad \text{and} \quad f(2, 30) = \tfrac{4}{18}.$$

We can use Definition 4.2.1 to compute the conditional pmf of Y given X for each of the possible values of X, $x = 0, 1, 2$. First, the marginal pmf of X is

$$f_X(0) = f(0,10) + f(0,20) = \frac{4}{18},$$

$$f_X(1) = f(1,10) + f(1,20) + f(1,30) = \frac{10}{18},$$

$$f_X(2) = f(2,30) = \frac{4}{18}.$$

For $x = 0$, $f(0,y)$ is positive only for $y = 10$ and $y = 20$. Thus $f(y|0)$ is positive only for $y = 10$ and $y = 20$, and

$$f(10|0) = \frac{f(0,10)}{f_X(0)} = \frac{\frac{2}{18}}{\frac{4}{18}} = \frac{1}{2},$$

$$f(20|0) = \frac{f(0,20)}{f_X(0)} = \frac{1}{2}.$$

That is, given the knowledge that $X = 0$, the conditional probability distribution for Y is the discrete distribution that assigns probability $\frac{1}{2}$ to each of the two points $y = 10$ and $y = 20$.

For $x = 1$, $f(y|1)$ is positive for $y = 10$, 20, and 30, and

$$f(10|1) = f(30|1) = \frac{\frac{3}{18}}{\frac{10}{18}} = \frac{3}{10},$$

$$f(20|1) = \frac{\frac{4}{18}}{\frac{10}{18}} = \frac{4}{10},$$

and for $x = 2$,

$$f(30|2) = \frac{\frac{4}{18}}{\frac{4}{18}} = 1.$$

The latter result reflects a fact that is also apparent from the joint pmf. If we know that $X = 2$, then we know that Y must be 30.

Other conditional probabilities can be computed using these conditional pmfs. For example,

$$P(Y > 10|X = 1) = f(20|1) + f(30|1) = \tfrac{7}{10}$$

or

$$P(Y > 10|X = 0) = f(20|0) = \tfrac{1}{2}. \qquad \|$$

If X and Y are continuous random variables, then $P(X = x) = 0$ for every value of x. To compute a conditional probability such as $P(Y > 200|X = 73)$, Definition 1.3.2 cannot be used since the denominator, $P(X = 73)$, is 0. Yet in actuality a value of X is observed. If, to the limit of our measurement, we see $X = 73$, this knowledge might give us information about Y (as the height and weight example at the beginning of this section indicated). It turns out that the appropriate way to define a conditional probability distribution for Y given $X = x$, when X and Y are both continuous, is analogous to the discrete case with pdfs replacing pmfs (see Miscellanea 4.9.3).

Definition 4.2.3 Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the *conditional pdf of Y given that $X = x$* is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $f_Y(y) > 0$, the *conditional pdf of X given that $Y = y$* is the function of x denoted by $f(x|y)$ and defined by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

To verify that $f(x|y)$ and $f(y|x)$ are indeed pdfs, the same steps can be used as in the earlier verification that Definition 4.2.1 had defined true pmfs with integrals now replacing sums.

In addition to their usefulness for calculating probabilities, the conditional pdfs or pmfs can also be used to calculate expected values. Just remember that $f(y|x)$ as a function of y is a pdf or pmf and use it in the same way that we have previously used unconditional pdfs or pmfs. If $g(Y)$ is a function of Y, then the *conditional expected value of $g(Y)$ given that $X = x$* is denoted by $\mathrm{E}(g(Y)|x)$ and is given by

$$\mathrm{E}(g(Y)|x) = \sum_y g(y) f(y|x) \quad \text{and} \quad \mathrm{E}(g(Y)|x) = \int_{-\infty}^{\infty} g(y) f(y|x)\, dy$$

in the discrete and continuous cases, respectively. The conditional expected value has all of the properties of the usual expected value listed in Theorem 2.2.5. Moreover, $\mathrm{E}(Y|X)$ provides the best guess at Y based on knowledge of X, extending the result in Example 2.2.6. (See Exercise 4.13.)

Example 4.2.4 (Calculating conditional pdfs) As in Example 4.1.12, let the continuous random vector (X, Y) have joint pdf $f(x, y) = e^{-y}$, $0 < x < y < \infty$. Suppose we wish to compute the conditional pdf of Y given $X = x$. The marginal pdf of X is computed as follows. If $x \leq 0$, $f(x, y) = 0$ for all values of y, so $f_X(x) = 0$. If $x > 0$, $f(x, y) > 0$ only if $y > x$. Thus

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

Thus, marginally, X has an exponential distribution. From Definition 4.2.3, the conditional distribution of Y given $X = x$ can be computed for any $x > 0$ (since these are the values for which $f_X(x) > 0$). For any such x,

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)}, \quad \text{if } y > x,$$

and

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = \frac{0}{e^{-x}} = 0, \quad \text{if } y \leq x.$$

Thus, given $X = x$, Y has an exponential distribution, where x is the location parameter in the distribution of Y and $\beta = 1$ is the scale parameter. The conditional distribution of Y is different for every value of x. It then follows that

$$\mathrm{E}(Y|X = x) = \int_x^\infty y e^{-(y-x)}\, dy = 1 + x.$$

The variance of the probability distribution described by $f(y|x)$ is called the *conditional variance of* Y *given* $X = x$. Using the notation $\mathrm{Var}(Y|x)$ for this, we have, using the ordinary definition of variance,

$$\mathrm{Var}(Y|x) = \mathrm{E}(Y^2|x) - (\mathrm{E}(Y|x))^2.$$

Applying this definition to our example, we obtain

$$\mathrm{Var}(Y|x) = \int_x^\infty y^2 e^{-(y-x)}\, dy - \left(\int_x^\infty y e^{-(y-x)}\, dy\right)^2 = 1.$$

In this case the conditional variance of Y given $X = x$ is the same for all values of x. In other situations, however, it may be different for different values of x. This conditional variance might be compared to the unconditional variance of Y. The marginal distribution of Y is gamma$(2, 1)$, which has $\mathrm{Var}\, Y = 2$. Given the knowledge that $X = x$, the variability in Y is considerably reduced. ||

A physical situation for which the model in Example 4.2.4 might be used is this. Suppose we have two light bulbs. The lengths of time each will burn are random variables denoted by X and Z. The lifelengths X and Z are independent and both have pdf $e^{-x}, x > 0$. The first bulb will be turned on. As soon as it burns out, the second bulb will be turned on. Now consider observing X, the time when the first bulb burns out, and $Y = X + Z$, the time when the second bulb burns out. Given that $X = x$ is when the first burned out and the second is started, $Y = Z + x$. This is like Example 3.5.3. The value x is acting as a location parameter, and the pdf of Y, in this case the *conditional* pdf of Y given $X = x$, is $f(y|x) = f_Z(y - x) = e^{-(y-x)}, y > x$.

The conditional distribution of Y given $X = x$ is possibly a different probability distribution for each value of x. Thus we really have a family of probability distributions for Y, one for each x. When we wish to describe this entire family, we will use the phrase "the distribution of $Y|X$." If, for example, X is a positive integer-valued random variable and the conditional distribution of Y given $X = x$ is binomial(x, p), then we might say the distribution of $Y|X$ is binomial(X, p) or write $Y|X \sim$ binomial(X, p). Whenever we use the symbol $Y|X$ or have a random variable as the parameter of a probability distribution, we are describing the family of conditional probability distributions. Joint pdfs or pmfs are sometimes defined by specifying the conditional $f(y|x)$ and the marginal $f_X(x)$. Then the definition yields $f(x, y) = f(y|x)f_X(x)$. These types of models are discussed more in Section 4.4.

Notice also that $\mathrm{E}(g(Y)|x)$ is a function of x. That is, for each value of x, $\mathrm{E}(g(Y)|x)$ is a real number obtained by computing the appropriate integral or sum. Thus, $\mathrm{E}(g(Y)|X)$ is a random variable whose value depends on the value of X. If $X = x$,

the value of the random variable $\mathrm{E}(g(Y)|X)$ is $\mathrm{E}(g(Y)|x)$. Thus, in Example 4.2.4, we can write $\mathrm{E}(Y|X) = 1 + X$.

In all the previous examples, the conditional distribution of Y given $X = x$ was different for different values of x. In some situations, the knowledge that $X = x$ does not give us any more information about Y than what we already had. This important relationship between X and Y is called *independence*. Just as with independent events in Chapter 1, it is more convenient to define independence in a symmetric fashion and then derive conditional properties like those we just mentioned. This we now do.

Definition 4.2.5 Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called *independent random variables* if, for every $x \in \Re$ and $y \in \Re$,

$$f(x, y) = f_X(x) f_Y(y). \tag{4.2.1}$$

If X and Y are independent, the conditional pdf of Y given $X = x$ is

$$\begin{aligned} f(y|x) &= \frac{f(x, y)}{f_X(x)} && \text{(definition)} \\ &= \frac{f_X(x) f_Y(y)}{f_X(x)} && \text{(from (4.2.1))} \\ &= f_Y(y), \end{aligned}$$

regardless of the value of x. Thus, for any $A \subset \Re$ and $x \in \Re$, $P(Y \in A|x) = \int_A f(y|x)\, dy = \int_A f_Y(y)\, dy = P(Y \in A)$. The knowledge that $X = x$ gives us no additional information about Y.

Definition 4.2.5 is used in two different ways. We might start with a joint pdf or pmf and then check whether X and Y are independent. To do this we must verify that (4.2.1) is true for every value of x and y. Or we might wish to define a model in which X and Y are independent. Consideration of what X and Y represent might indicate that knowledge that $X = x$ should give us no information about Y. In this case we could specify the marginal distributions of X and Y and then define the joint distribution as the product given in (4.2.1).

Example 4.2.6 (Checking independence–I) Consider the discrete bivariate random vector (X, Y), with joint pmf given by

$$f(10, 1) = f(20, 1) = f(20, 2) = \tfrac{1}{10},$$
$$f(10, 2) = f(10, 3) = \tfrac{1}{5}, \quad \text{and} \quad f(20, 3) = \tfrac{3}{10}.$$

The marginal pmfs are easily calculated to be

$$f_X(10) = f_X(20) = \frac{1}{2} \quad \text{and} \quad f_Y(1) = \frac{1}{5},\ f_Y(2) = \frac{3}{10},\ \text{and}\ f_Y(3) = \frac{1}{2}.$$

The random variables X and Y are not independent because (4.2.1) is not true for every x and y. For example,

$$f(10, 3) = \frac{1}{5} \neq \frac{1}{2}\frac{1}{2} = f_X(10) f_Y(3).$$

The relationship (4.2.1) must hold for every choice of x and y if X and Y are to be independent. Note that $f(10,1) = \frac{1}{10} = \frac{1}{2}\frac{1}{5} = f_X(10)f_Y(1)$. That (4.2.1) holds for *some* values of x and y does not ensure that X and Y are independent. All values must be checked. ||

The verification that X and Y are independent by direct use of (4.2.1) would require the knowledge of $f_X(x)$ and $f_Y(y)$. The following lemma makes the verification somewhat easier.

Lemma 4.2.7 *Let (X,Y) be a bivariate random vector with joint pdf or pmf $f(x,y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \Re$ and $y \in \Re$,*

$$f(x,y) = g(x)h(y).$$

Proof: The "only if" part is proved by defining $g(x) = f_X(x)$ and $h(y) = f_Y(y)$ and using (4.2.1). To prove the "if" part for continuous random variables, suppose that $f(x,y) = g(x)h(y)$. Define

$$\int_{-\infty}^{\infty} g(x)\,dx = c \quad \text{and} \quad \int_{-\infty}^{\infty} h(y)\,dy = d,$$

where the constants c and d satisfy

$$\begin{aligned} cd &= \left(\int_{-\infty}^{\infty} g(x)\,dx\right)\left(\int_{-\infty}^{\infty} h(y)\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\,dx\,dy \qquad (4.2.2) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy \\ &= 1. \qquad (f(x,y) \text{ is a joint pdf}) \end{aligned}$$

Furthermore, the marginal pdfs are given by

(4.2.3)

$$f_X(x) = \int_{-\infty}^{\infty} g(x)h(y)\,dy = g(x)d \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} g(x)h(y)\,dx = h(y)c.$$

Thus, using (4.2.2) and (4.2.3), we have

$$f(x,y) = g(x)h(y) = g(x)h(y)cd = f_X(x)f_Y(y),$$

showing that X and Y are independent. Replacing integrals with sums proves the lemma for discrete random vectors. □

Example 4.2.8 (Checking independence–II) Consider the joint pdf $f(x,y) = \frac{1}{384}x^2y^4e^{-y-(x/2)}$, $x > 0$ and $y > 0$. If we define

$$g(x) = \begin{cases} x^2e^{-x/2} & x > 0 \\ 0 & x \le 0 \end{cases} \quad \text{and} \quad h(y) = \begin{cases} y^4e^{-y}/384 & y > 0 \\ 0 & y \le 0, \end{cases}$$

then $f(x,y) = g(x)h(y)$ for all $x \in \Re$ and all $y \in \Re$. By Lemma 4.2.7, we conclude that X and Y are independent random variables. We do not have to compute marginal pdfs. ||

If X and Y are independent random variables, then from (4.2.1) it is clear that $f(x,y) > 0$ on the set $\{(x,y) : x \in A \text{ and } y \in B\}$, where $A = \{x : f_X(x) > 0\}$ and $B = \{y : f_Y(y) > 0\}$. A set of this form is called a cross-product and is usually denoted by $A \times B$. Membership in a cross-product can be checked by considering the x and y values separately. If $f(x,y)$ is a joint pdf or pmf and the set where $f(x,y) > 0$ is not a cross-product, then the random variables X and Y with joint pdf or pmf $f(x,y)$ are not independent. In Example 4.2.4, the set $0 < x < y < \infty$ is not a cross-product. To check membership in this set we must check that not only $0 < x < \infty$ and $0 < y < \infty$ but also $x < y$. Thus the random variables in Example 4.2.4 are not independent. Example 4.2.2 gives an example of a joint pmf that is positive on a set that is not a cross-product.

Example 4.2.9 (Joint probability model) As an example of using independence to define a joint probability model, consider this situation. A student from an elementary school in Kansas City is randomly selected and X = the number of living parents of the student is recorded. Suppose the marginal distribution of X is

$$f_X(0) = .01, \quad f_X(1) = .09, \quad \text{and} \quad f_X(2) = .90.$$

A retiree from Sun City is randomly selected and Y = the number of living parents of the retiree is recorded. Suppose the marginal distribution of Y is

$$f_Y(0) = .70, \quad f_Y(1) = .25, \quad \text{and} \quad f_Y(2) = .05.$$

It seems reasonable to assume that these two random variables are independent. Knowledge of the number of parents of the student tells us nothing about the number of parents of the retiree. The only joint distribution of X and Y that reflects this independence is the one defined by (4.2.1). Thus, for example,

$$f(0,0) = f_X(0)f_Y(0) = .0070 \quad \text{and} \quad f(0,1) = f_X(0)f_Y(1) = .0025.$$

This joint distribution can be used to calculate quantities such as

$$\begin{aligned} P(X = Y) &= f(0,0) + f(1,1) + f(2,2) \\ &= (.01)(.70) + (.09)(.25) + (.90)(.05) = .0745. \end{aligned}$$ ||

Certain probabilities and expectations are easy to calculate if X and Y are independent, as the next theorem indicates.

Theorem 4.2.10 *Let X and Y be independent random variables.*

a. *For any $A \subset \Re$ and $B \subset \Re$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.*

b. *Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then*

$$\mathrm{E}\left(g(X)h(Y)\right) = \left(\mathrm{E}g(X)\right)\left(Eh(Y)\right).$$

Proof: For continuous random variables, part (b) is proved by noting that

$$\begin{aligned}
\mathrm{E}\left(g(X)h(Y)\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,dx\,dy \qquad \text{(by (4.2.1))} \\
&= \int_{-\infty}^{\infty} h(y)f_Y(y)\int_{-\infty}^{\infty} g(x)f_X(x)\,dx\,dy \\
&= \left(\int_{-\infty}^{\infty} g(x)f_X(x)\,dx\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)\,dy\right) \\
&= \left(\mathrm{E}g(X)\right)\left(\mathrm{E}h(Y)\right).
\end{aligned}$$

The result for discrete random variables is proved by replacing integrals by sums. Part (a) can be proved by a series of steps similar to those above or by the following argument. Let $g(x)$ be the indicator function of the set A. Let $h(y)$ be the indicator function of the set B. Note that $g(x)h(y)$ is the indicator function of the set $C \subset \Re^2$ defined by $C = \{(x,y) : x \in A, y \in B\}$. Also note that for an indicator function such as $g(x), \mathrm{E}g(X) = P(X \in A)$. Thus using the expectation equality just proved, we have

$$\begin{aligned}
P(X \in A, Y \in B) &= P\left((X,Y) \in C\right) = \mathrm{E}\left(g(X)h(Y)\right) \\
&= \left(\mathrm{E}g(X)\right)\left(\mathrm{E}h(Y)\right) = P(X \in A)P(Y \in B).
\end{aligned}$$

□

Example 4.2.11 (Expectations of independent variables) Let X and Y be independent exponential(1) random variables. From Theorem 4.2.10 we have

$$P(X \geq 4, Y < 3) = P(X \geq 4)P(Y < 3) = e^{-4}(1 - e^{-3}).$$

Letting $g(x) = x^2$ and $h(y) = y$, we see that

$$\mathrm{E}\left(X^2Y\right) = \left(\mathrm{E}X^2\right)\left(\mathrm{E}Y\right) = \left(\mathrm{Var}\,X + (\mathrm{E}X)^2\right)\mathrm{E}Y = (1 + 1^2)1 = 2. \qquad \|$$

The following result concerning sums of independent random variables is a simple consequence of Theorem 4.2.10.

Theorem 4.2.12 *Let X and Y be independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then the moment generating function of the random variable $Z = X + Y$ is given by*

$$M_Z(t) = M_X(t)M_Y(t).$$

Proof: Using the definition of the mgf and Theorem 4.2.10, we have

$$M_Z(t) = \mathrm{E}e^{tZ} = \mathrm{E}e^{t(X+Y)} = \mathrm{E}(e^{tX}e^{tY}) = (\mathrm{E}e^{tX})(\mathrm{E}e^{tY}) = M_X(t)M_Y(t). \qquad \square$$

Example 4.2.13 (Mgf of a sum of normal variables) Sometimes Theorem 4.2.12 can be used to easily derive the distribution of Z from knowledge of the distribution of X and Y. For example, let $X \sim \mathrm{n}(\mu, \sigma^2)$ and $Y \sim \mathrm{n}(\gamma, \tau^2)$ be independent normal random variables. From Exercise 2.33, the mgfs of X and Y are

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2) \quad \text{and} \quad M_Y(t) = \exp(\gamma t + \tau^2 t^2/2).$$

Thus, from Theorem 4.2.12, the mgf of $Z = X + Y$ is

$$M_Z(t) = M_X(t) M_Y(t) = \exp\left((\mu + \gamma)t + (\sigma^2 + \tau^2)t^2/2\right).$$

This is the mgf of a normal random variable with mean $\mu + \gamma$ and variance $\sigma^2 + \tau^2$. This result is important enough to be stated as a theorem. ||

Theorem 4.2.14 *Let $X \sim \mathrm{n}(\mu, \sigma^2)$ and $Y \sim \mathrm{n}(\gamma, \tau^2)$ be independent normal random variables. Then the random variable $Z = X + Y$ has a $\mathrm{n}(\mu + \gamma, \sigma^2 + \tau^2)$ distribution.*

If $f(x, y)$ is the joint pdf for the continuous random vector (X, Y), (4.2.1) may fail to hold on a set A of (x, y) values for which $\int_A \int dx\, dy = 0$. In such a case X and Y are still called independent random variables. This reflects the fact that two pdfs that differ only on a set such as A define the same probability distribution for (X, Y). To see this, suppose $f(x, y)$ and $f^*(x, y)$ are two pdfs that are equal everywhere except on a set A for which $\int_A \int dx\, dy = 0$. Let (X, Y) have pdf $f(x, y)$, let (X^*, Y^*) have pdf $f^*(x, y)$, and let B be any subset of $\Re^2$. Then

$$\begin{aligned} P((X, Y) \in B) &= \int_B \int f(x, y)\, dx\, dy \\ &= \int_{B \cap A^c} \int f(x, y)\, dx\, dy \\ &= \int_{B \cap A^c} \int f^*(x, y)\, dx\, dy \\ &= \int_B \int f^*(x, y)\, dx\, dy = P((X^*, Y^*) \in B). \end{aligned}$$

Thus (X, Y) and (X^*, Y^*) have the same probability distribution. So, for example, $f(x, y) = e^{-x-y}$, $x > 0$ and $y > 0$, is a pdf for two independent exponential random variables and satisfies (4.2.1). But, $f^*(x, y)$, which is equal to $f(x, y)$ except that $f^*(x, y) = 0$ if $x = y$, is also the pdf for two independent exponential random variables even though (4.2.1) is not true on the set $A = \{(x, x) : x > 0\}$.

4.3 Bivariate Transformations

In Section 2.1, methods of finding the distribution of a function of a random variable were discussed. In this section we extend these ideas to the case of bivariate random vectors.

Let (X, Y) be a bivariate random vector with a known probability distribution. Now consider a new bivariate random vector (U, V) defined by $U = g_1(X, Y)$ and $V =$

$g_2(X,Y)$, where $g_1(x,y)$ and $g_2(x,y)$ are some specified functions. If B is any subset of $\Re^2$, then $(U,V) \in B$ if and only if $(X,Y) \in A$, where $A = \{(x,y) : (g_1(x,y), g_2(x,y)) \in B\}$. Thus $P((U,V) \in B) = P((X,Y) \in A)$, and the probability distribution of (U,V) is completely determined by the probability distribution of (X,Y).

If (X,Y) is a discrete bivariate random vector, then there is only a countable set of values for which the joint pmf of (X,Y) is positive. Call this set $\mathcal{A}$. Define the set $\mathcal{B} = \{(u,v) : u = g_1(x,y) \text{ and } v = g_2(x,y) \text{ for some } (x,y) \in \mathcal{A}\}$. Then $\mathcal{B}$ is the countable set of possible values for the discrete random vector (U,V). And if, for any $(u,v) \in \mathcal{B}$, A_{uv} is defined to be $\{(x,y) \in \mathcal{A} : g_1(x,y) = u \text{ and } g_2(x,y) = v\}$, then the joint pmf of (U,V), $f_{U,V}(u,v)$, can be computed from the joint pmf of (X,Y) by

$$(4.3.1)\quad f_{U,V}(u,v) = P(U=u, V=v) = P((X,Y) \in A_{uv}) = \sum_{(x,y)\in A_{uv}} f_{X,Y}(x,y).$$

Example 4.3.1 (Distribution of the sum of Poisson variables) Let X and Y be independent Poisson random variables with parameters θ and λ, respectively. Thus the joint pmf of (X,Y) is

$$f_{X,Y}(x,y) = \frac{\theta^x e^{-\theta}}{x!}\frac{\lambda^y e^{-\lambda}}{y!}, \quad x = 0,1,2,\ldots,\ y = 0,1,2,\ldots.$$

The set $\mathcal{A}$ is $\{(x,y) : x = 0,1,2,\ldots \text{ and } y = 0,1,2,\ldots\}$. Now define $U = X + Y$ and $V = Y$. That is, $g_1(x,y) = x + y$ and $g_2(x,y) = y$. We will describe the set $\mathcal{B}$, the set of possible (u,v) values. The possible values for v are the nonnegative integers. The variable $v = y$ and thus has the same set of possible values. For a given value of v, $u = x + y = x + v$ must be an integer greater than or equal to v since x is a nonnegative integer. The set of all possible (u,v) values is thus given by $\mathcal{B} = \{(u,v) : v = 0,1,2,\ldots \text{ and } u = v, v+1, v+2,\ldots\}$. For any $(u,v) \in \mathcal{B}$, the only (x,y) value satisfying $x + y = u$ and $y = v$ is $x = u - v$ and $y = v$. Thus, in this example, A_{uv} always consists of only the single point $(u-v, v)$. From (4.3.1) we thus obtain the joint pmf of (U,V) as

$$f_{U,V}(u,v) = f_{X,Y}(u-v,v) = \frac{\theta^{u-v}e^{-\theta}}{(u-v)!}\frac{\lambda^v e^{-\lambda}}{v!}, \quad \begin{array}{l} v = 0,1,2,\ldots, \\ u = v, v+1, v+2,\ldots. \end{array}$$

In this example it is interesting to compute the marginal pmf of U. For any fixed nonnegative integer u, $f_{U,V}(u,v) > 0$ only for $v = 0,1,\ldots,u$. This gives the set of v values to sum over to obtain the marginal pmf of U. It is

$$f_U(u) = \sum_{v=0}^{u} \frac{\theta^{u-v}e^{-\theta}}{(u-v)!}\frac{\lambda^v e^{-\lambda}}{v!} = e^{-(\theta+\lambda)}\sum_{v=0}^{u}\frac{\theta^{u-v}}{(u-v)!}\frac{\lambda^v}{v!}, \quad u = 0,1,2,\ldots.$$

This can be simplified by noting that, if we multiply and divide each term by $u!$, we can use the Binomial Theorem to obtain

$$f_U(u) = \frac{e^{-(\theta+\lambda)}}{u!}\sum_{v=0}^{u}\binom{u}{v}\lambda^v\theta^{u-v} = \frac{e^{-(\theta+\lambda)}}{u!}(\theta+\lambda)^u, \quad u = 0,1,2,\ldots.$$

This is the pmf of a Poisson random variable with parameter $\theta + \lambda$. This result is significant enough to be stated as a theorem. ||

Theorem 4.3.2 *If $X \sim \text{Poisson}(\theta)$ and $Y \sim \text{Poisson}(\lambda)$ and X and Y are independent, then $X + Y \sim \text{Poisson}(\theta + \lambda)$.*

If (X, Y) is a continuous random vector with joint pdf $f_{X,Y}(x, y)$, then the joint pdf of (U, V) can be expressed in terms of $f_{X,Y}(x, y)$ in a manner analogous to (2.1.8). As before, $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$ and $\mathcal{B} = \{(u, v) : u = g_1(x, y) \text{ and } v = g_2(x, y) \text{ for some } (x, y) \in \mathcal{A}\}$. The joint pdf $f_{U,V}(u, v)$ will be positive on the set $\mathcal{B}$. For the simplest version of this result we assume that the transformation $u = g_1(x, y)$ and $v = g_2(x, y)$ defines a one-to-one transformation of $\mathcal{A}$ onto $\mathcal{B}$. The transformation is onto because of the definition of $\mathcal{B}$. We are assuming that for each $(u, v) \in \mathcal{B}$ there is only one $(x, y) \in \mathcal{A}$ such that $(u, v) = (g_1(x, y), g_2(x, y))$. For such a one-to-one, onto transformation, we can solve the equations $u = g_1(x, y)$ and $v = g_2(x, y)$ for x and y in terms of u and v. We will denote this inverse transformation by $x = h_1(u, v)$ and $y = h_2(u, v)$. The role played by a derivative in the univariate case is now played by a quantity called the *Jacobian of the transformation*. This function of (u, v), denoted by J, is the *determinant of a matrix* of partial derivatives. It is defined by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v},$$

where

$$\frac{\partial x}{\partial u} = \frac{\partial h_1(u, v)}{\partial u}, \quad \frac{\partial x}{\partial v} = \frac{\partial h_1(u, v)}{\partial v}, \quad \frac{\partial y}{\partial u} = \frac{\partial h_2(u, v)}{\partial u}, \quad \text{and} \quad \frac{\partial y}{\partial v} = \frac{\partial h_2(u, v)}{\partial v}.$$

We assume that J is not identically 0 on $\mathcal{B}$. Then the joint pdf of (U, V) is 0 outside the set $\mathcal{B}$ and on the set $\mathcal{B}$ is given by

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v),\ h_2(u, v))|J|, \tag{4.3.2}$$

where $|J|$ is the absolute value of J. When we use (4.3.2), it is sometimes just as difficult to determine the set $\mathcal{B}$ and verify that the transformation is one-to-one as it is to substitute into formula (4.3.2). Note these parts of the explanations in the following examples.

Example 4.3.3 (Distribution of the product of beta variables) Let $X \sim \text{beta}(\alpha, \beta)$ and $Y \sim \text{beta}(\alpha + \beta, \gamma)$ be independent random variables. The joint pdf of (X, Y) is

$$f_{X,Y}(x, y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \frac{\Gamma(\alpha + \beta + \gamma)}{\Gamma(\alpha + \beta)\Gamma(\gamma)} y^{\alpha+\beta-1}(1 - y)^{\gamma-1},$$
$$0 < x < 1, \quad 0 < y < 1.$$

Consider the transformation $U = XY$ and $V = X$. The set of possible values for V is $0 < v < 1$ since $V = X$. For a fixed value of $V = v$, U must be between 0 and v since $X = V = v$ and Y is between 0 and 1. Thus, this transformation maps the set

$\mathcal{A}$ onto the set $\mathcal{B} = \{(u,v) : 0 < u < v < 1\}$. For any $(u,v) \in \mathcal{B}$, the equations $u = xy$ and $v = x$ can be uniquely solved for $x = h_1(u,v) = v$ and $y = h_2(u,v) = u/v$. Note that if considered as a transformation defined on all of $\Re^2$, this transformation is not one-to-one. Any point $(0,y)$ is mapped into the point $(0,0)$. But as a function defined only on $\mathcal{A}$, it is a one-to-one transformation onto $\mathcal{B}$. The Jacobian is given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}.$$

Thus, from (4.3.2) we obtain the joint pdf as

$$f_{U,V}(u,v) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} v^{\alpha-1}(1-v)^{\beta-1}\left(\frac{u}{v}\right)^{\alpha+\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\frac{1}{v}, \quad 0 < u < v < 1. \tag{4.3.3}$$

The marginal distribution of $V = X$ is, of course, a beta(α,β) distribution. But the distribution of U is also a beta distribution:

$$\begin{aligned} f_U(u) &= \int_u^1 f_{U,V}(u,v)\,dv \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}\int_u^1 \left(\frac{u}{v}-u\right)^{\beta-1}\left(1-\frac{u}{v}\right)^{\gamma-1}\left(\frac{u}{v^2}\right)dv. \end{aligned}$$

The expression (4.3.3) was used but some terms have been rearranged. Now make the univariate change of variable $y = (u/v - u)/(1-u)$ so that $dy = -u/[v^2(1-u)]dv$ to obtain

$$\begin{aligned} f_U(u) &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}\int_0^1 y^{\beta-1}(1-y)^{\gamma-1}\,dy \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}\frac{\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\beta+\gamma)} \\ &= \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta+\gamma)} u^{\alpha-1}(1-u)^{\beta+\gamma-1}, \quad 0 < u < 1. \end{aligned}$$

To obtain the second identity we recognized the integrand as the kernel of a beta pdf and used (3.3.17). Thus we see that the marginal distribution of U is beta$(\alpha, \beta+\gamma)$. ||

Example 4.3.4 (Sum and difference of normal variables) Let X and Y be independent, standard normal random variables. Consider the transformation $U = X+Y$ and $V = X-Y$. In the notation used above, $U = g_1(X,Y)$ where $g_1(x,y) = x+y$ and $V = g_2(X,Y)$ where $g_2(x,y) = x-y$. The joint pdf of X and Y is, of course, $f_{X,Y}(x,y) = (2\pi)^{-1}\exp(-x^2/2)\exp(-y^2/2)$, $-\infty < x < \infty$, $-\infty < y < \infty$. So the set

$\mathcal{A} = \Re^2$. To determine the set $\mathcal{B}$ on which $f_{U,V}(u,v)$ is positive, we must determine all the values that

$$u = x + y \quad \text{and} \quad v = x - y \tag{4.3.4}$$

take on as (x,y) range over the set $\mathcal{A} = \Re^2$. But we can set u to be any number and v to be any number and uniquely solve equations (4.3.4) for x and y to obtain

$$x = h_1(u,v) = \frac{u+v}{2} \quad \text{and} \quad y = h_2(u,v) = \frac{u-v}{2}. \tag{4.3.5}$$

This shows two things. For any $(u,v) \in \Re^2$ there is an $(x,y) \in \mathcal{A}$ (defined by (4.3.5)) such that $u = x + y$ and $v = x - y$. So $\mathcal{B}$, the set of all possible (u,v) values, is $\Re^2$. Since the solution (4.3.5) is unique, this also shows that the transformation we have considered is one-to-one. Only the (x,y) given in (4.3.5) will yield $u = x + y$ and $v = x - y$. From (4.3.5) the partial derivatives of x and y are easy to compute. We obtain

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Substituting the expressions (4.3.5) for x and y into $f_{X,Y}(x,y)$ and using $|J| = \frac{1}{2}$, we obtain the joint pdf of (U,V) from (4.3.2) as

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v))|J| = \frac{1}{2\pi} e^{-((u+v)/2)^2/2} e^{-((u-v)/2)^2/2} \frac{1}{2}$$

for $-\infty < u < \infty$ and $-\infty < v < \infty$. Multiplying out the squares in the exponentials, we see that the terms involving uv cancel. Thus after some simplification and rearrangement we obtain

$$f_{U,V}(u,v) = \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-u^2/4} \right) \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-v^2/4} \right).$$

The joint pdf has factored into a function of u and a function of v. By Lemma 4.2.7, U and V are independent. From Theorem 4.2.14, the marginal distribution of $U = X+Y$ is $\mathrm{n}(0,2)$. Similarly, Theorem 4.2.12 could be used to find that the marginal distribution of V is also $\mathrm{n}(0,2)$. This important fact, that sums and differences of independent normal random variables are independent normal random variables, is true regardless of the means of X and Y, so long as $\mathrm{Var}\, X = \mathrm{Var}\, Y$. This result is left as Exercise 4.27. Theorems 4.2.12 and 4.2.14 give us the marginal distributions of U and V. But the more involved analysis here is required to determine that U and V are independent. ||

In Example 4.3.4, we found that U and V are independent random variables. There is a much simpler, but very important, situation in which new variables U and V, defined in terms of original variables X and Y, are independent. Theorem 4.3.5 describes this.

Theorem 4.3.5 *Let X and Y be independent random variables. Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then the random variables $U = g(X)$ and $V = h(Y)$ are independent.*

Proof: We will prove the theorem assuming U and V are continuous random variables. For any $u \in \Re$ and $v \in \Re$, define

$$A_u = \{x : g(x) \le u\} \quad \text{and} \quad B_v = \{y : h(y) \le v\}.$$

Then the joint cdf of (U, V) is

$$\begin{aligned} F_{U,V}(u,v) &= P(U \le u, V \le v) && \text{(definition of cdf)} \\ &= P(X \in A_u, Y \in B_v) && \text{(definition of } U \text{ and } V) \\ &= P(X \in A_u)P(Y \in B_v). && \text{(Theorem 4.2.10)} \end{aligned}$$

The joint pdf of (U, V) is

$$\begin{aligned} f_{U,V}(u,v) &= \frac{\partial^2}{\partial u \partial v} F_{U,V}(u,v) && \text{(by (4.1.4))} \\ &= \left(\frac{d}{du} P(X \in A_u)\right)\left(\frac{d}{dv} P(Y \in B_v)\right), \end{aligned}$$

where, as the notation indicates, the first factor is a function only of u and the second factor is a function only of v. Hence, by Lemma 4.2.7, U and V are independent. □

It may be that there is only one function, say $U = g_1(X, Y)$, of interest. In such cases, this method may still be used to find the distribution of U. If another convenient function, $V = g_2(X, Y)$, can be chosen so that the resulting transformation from (X, Y) to (U, V) is one-to-one on $\mathcal{A}$, then the joint pdf of (U, V) can be derived using (4.3.2) and the marginal pdf of U can be obtained from the joint pdf. In the previous example, perhaps we were interested only in $U = XY$. We could choose to define $V = X$, recognizing that the resulting transformation is one-to-one on $\mathcal{A}$. Then we would proceed as in the example to obtain the marginal pdf of U. But other choices, such as $V = Y$, would work as well (see Exercise 4.23).

Of course, in many situations, the transformation of interest is not one-to-one. Just as Theorem 2.1.8 generalized the univariate method to many-to-one functions, the same can be done here. As before, $\mathcal{A} = \{(x, y) : f_{X,Y}(x, y) > 0\}$. Suppose $A_0, A_1, \ldots, A_k$ form a partition of $\mathcal{A}$ with these properties. The set A_0, which may be empty, satisfies $P((X, Y) \in A_0) = 0$. The transformation $U = g_1(X, Y)$ and $V = g_2(X, Y)$ is a one-to-one transformation from A_i onto $\mathcal{B}$ for each $i = 1, 2, \ldots, k$. Then for each i, the inverse functions from $\mathcal{B}$ to A_i can be found. Denote the ith inverse by $x = h_{1i}(u, v)$ and $y = h_{2i}(u, v)$. This ith inverse gives, for $(u, v) \in \mathcal{B}$, the unique $(x, y) \in A_i$ such that $(u, v) = (g_1(x, y), g_2(x, y))$. Let J_i denote the Jacobian computed from the ith inverse. Then assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the joint pdf, $f_{U,V}(u, v)$:

$$f_{U,V}(u,v) = \sum_{i=1}^{k} f_{X,Y}(h_{1i}(u,v), h_{2i}(u,v))|J_i|. \tag{4.3.6}$$

Example 4.3.6 (Distribution of the ratio of normal variables) Let X and Y be independent n$(0,1)$ random variables. Consider the transformation $U = X/Y$ and $V = |Y|$. (U and V can be defined to be any value, say $(1,1)$, if $Y = 0$ since $P(Y = 0) = 0$.) This transformation is not one-to-one since the points (x,y) and $(-x,-y)$ are both mapped into the same (u,v) point. But if we restrict consideration to either positive or negative values of y, then the transformation is one-to-one. In the above notation, let

$$A_1 = \{(x,y) : y > 0\}, \quad A_2 = \{(x,y) : y < 0\}, \quad \text{and} \quad A_0 = \{(x,y) : y = 0\}.$$

A_0, A_1, and A_2 form a partition of $\mathcal{A} = \Re^2$ and $P((X,Y) \in A_0) = P(Y = 0) = 0$. For either A_1 or A_2, if $(x,y) \in A_i$, $v = |y| > 0$, and for a fixed value of $v = |y|$, $u = x/y$ can be any real number since x can be any real number. Thus, $\mathcal{B} = \{(u,v) : v > 0\}$ is the image of both A_1 and A_2 under the transformation. Furthermore, the inverse transformations from $\mathcal{B}$ to A_1 and $\mathcal{B}$ to A_2 are given by $x = h_{11}(u,v) = uv$, $y = h_{21}(u,v) = v$, and $x = h_{12}(u,v) = -uv$, $y = h_{22}(u,v) = -v$. Note that the first inverse gives positive values of y and the second gives negative values of y. The Jacobians from the two inverses are $J_1 = J_2 = v$. Using

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-x^2/2} e^{-y^2/2},$$

from (4.3.6) we obtain

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{2\pi} e^{-(uv)^2/2} e^{-v^2/2} |v| + \frac{1}{2\pi} e^{-(-uv)^2/2} e^{-(-v)^2/2} |v| \\ &= \frac{v}{\pi} e^{-(u^2+1)v^2/2}, \quad -\infty < u < \infty, \quad 0 < v < \infty. \end{aligned}$$

From this the marginal pdf of U can be computed to be

$$\begin{aligned} f_U(u) &= \int_0^\infty \frac{v}{\pi} e^{-(u^2+1)v^2/2}\, dv \\ &= \frac{1}{2\pi} \int_0^\infty e^{-(u^2+1)z/2}\, dz \quad z = v^2) \qquad \text{(change of variable)} \\ &= \frac{1}{2\pi} \frac{2}{(u^2+1)} \qquad \left(\begin{array}{c}\text{integrand is kernel of} \\ \text{exponential } (\beta = 2/(u^2+1)) \text{ pdf}\end{array}\right) \\ &= \frac{1}{\pi(u^2+1)}, \quad -\infty < u < \infty. \end{aligned}$$

So we see that the ratio of two independent standard normal random variables is a Cauchy random variable. (See Exercise 4.28 for more relationships between normal and Cauchy random variables.) ||

4.4 Hierarchical Models and Mixture Distributions

In the cases we have seen thus far, a random variable has a single distribution, possibly depending on parameters. While, in general, a random variable can have only one distribution, it is often easier to model a situation by thinking of things in a hierarchy.

Example 4.4.1 (Binomial-Poisson hierarchy) Perhaps the most classic hierarchical model is the following. An insect lays a large number of eggs, each surviving with probability p. On the average, how many eggs will survive?

The "large number" of eggs laid is a random variable, often taken to be Poisson(λ). Furthermore, if we assume that each egg's survival is independent, then we have Bernoulli trials. Therefore, if we let X = number of survivors and Y = number of eggs laid, we have

$$X|Y \sim \text{binomial}(Y, p),$$
$$Y \sim \text{Poisson}(\lambda),$$

a hierarchical model. (Recall that we use notation such as $X|Y \sim \text{binomial}(Y,p)$ to mean that the conditional distribution of X given $Y = y$ is binomial(y,p).) ||

The advantage of the hierarchy is that complicated processes may be modeled by a sequence of relatively simple models placed in a hierarchy. Also, dealing with the hierarchy is no more difficult than dealing with conditional and marginal distributions.

Example 4.4.2 (Continuation of Example 4.4.1) The random variable of interest, X = number of survivors, has the distribution given by

$$\begin{aligned} P(X=x) &= \sum_{y=0}^{\infty} P(X=x, Y=y) \\ &= \sum_{y=0}^{\infty} P(X=x|Y=y)P(Y=y) \qquad \left(\begin{matrix}\text{definition of}\\ \text{conditional probability}\end{matrix}\right) \\ &= \sum_{y=x}^{\infty} \left[\binom{y}{x} p^x (1-p)^{y-x}\right]\left[\frac{e^{-\lambda}\lambda^y}{y!}\right], \quad \left(\begin{matrix}\text{conditional probability}\\ \text{is 0 if } y<x\end{matrix}\right) \end{aligned}$$

since $X|Y = y$ is binomial(y,p) and Y is Poisson(λ). If we now simplify this last expression, canceling what we can and multiplying by λ^x/λ^x, we get

$$\begin{aligned} P(X=x) &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{((1-p)\lambda)^t}{t!} \quad (t = y-x) \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} \qquad \left(\begin{matrix}\text{sum is a kernel for}\\ \text{a Poisson distribution}\end{matrix}\right) \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p}, \end{aligned}$$

so $X \sim$ Poisson(λp). Thus, any marginal inference on X is with respect to a Poisson(λp) distribution, with Y playing no part at all. Introducing Y in the hierarchy

was mainly to aid our understanding of the model. There was an added bonus in that the parameter of the distribution of X is the product of two parameters, each relatively simple to understand.

The answer to the original question is now easy to compute:

$$\mathrm{E}X = \lambda p,$$

so, on the average, λp eggs will survive. If we were interested only in this mean and did not need the distribution, we could have used properties of conditional expectations. ‖

Sometimes, calculations can be greatly simplified be using the following theorem. Recall from Section 4.2 that $\mathrm{E}(X|y)$ is a function of y and $\mathrm{E}(X|Y)$ is a random variable whose value depends on the value of Y.

Theorem 4.4.3 *If X and Y are any two random variables, then*

$$\mathrm{E}X = \mathrm{E}\left(\mathrm{E}(X|Y)\right), \tag{4.4.1}$$

provided that the expectations exist.

Proof: Let $f(x,y)$ denote the joint pdf of X and Y. By definition, we have

$$\mathrm{E}X = \int\int x f(x,y)\,dx\,dy = \int \left[\int x f(x|y)\,dx\right] f_Y(y)\,dy, \tag{4.4.2}$$

where $f(x|y)$ and $f_Y(y)$ are the conditional pdf of X given $Y = y$ and the marginal pdf of Y, respectively. But now notice that the inner integral in (4.4.2) is the conditional expectation $\mathrm{E}(X|y)$, and we have

$$\mathrm{E}X = \int \mathrm{E}(X|y) f_Y(y)\,dy = \mathrm{E}\left(\mathrm{E}(X|Y)\right),$$

as desired. Replace integrals by sums to prove the discrete case. □

Note that equation (4.4.1) contains an abuse of notation, since we have used the "E" to stand for different expectations in the same equation. The "E" in the left-hand side of (4.4.1) is expectation with respect to the marginal distribution of X. The first "E" in the right-hand side of (4.4.1) is expectation with respect to the marginal distribution of Y, while the second one stands for expectation with respect to the conditional distribution of $X|Y$. However, there is really no cause for confusion because these interpretations are the only ones that the symbol "E" can take!

We can now easily compute the expected number of survivors in Example 4.4.1. From Theorem 4.4.3 we have

$$\begin{aligned}
\mathrm{E}X &= \mathrm{E}\left(\mathrm{E}(X|Y)\right) \\
&= \mathrm{E}(pY) && (\text{since } X|Y \sim \text{binomial}(Y,p)) \\
&= p\lambda. && (\text{since } Y \sim \text{Poisson}(\lambda))
\end{aligned}$$

The term *mixture distribution* in the title of this section refers to a distribution arising from a hierarchical structure. Although there is no standardized definition for this term, we will use the following definition, which seems to be a popular one.

Definition 4.4.4 A random variable X is said to have a *mixture distribution* if the distribution of X depends on a quantity that also has a distribution.

Thus, in Example 4.4.1 the Poisson(λp) distribution is a mixture distribution since it is the result of combining a binomial(Y, p) with $Y \sim$ Poisson(λ). In general, we can say that hierarchical models lead to mixture distributions.

There is nothing to stop the hierarchy at two stages, but it should be easy to see that any more complicated hierarchy can be treated as a two-stage hierarchy theoretically. There may be advantages, however, in modeling a phenomenon as a multistage hierarchy. It may be easier to understand.

Example 4.4.5 (Generalization of Example 4.4.1) Consider a generalization of Example 4.4.1, where instead of one mother insect there are a large number of mothers and one mother is chosen at random. We are still interested in knowing the average number of survivors, but it is no longer clear that the number of eggs laid follows the same Poisson distribution for each mother. The following three-stage hierarchy may be more appropriate. Let X = number of survivors in a litter; then

$$\begin{aligned} X|Y &\sim \text{binomial}(Y, p), \\ Y|\Lambda &\sim \text{Poisson}(\Lambda), \\ \Lambda &\sim \text{exponential}(\beta), \end{aligned}$$

where the last stage of the hierarchy accounts for the variability across different mothers.

The mean of X can easily be calculated as

$$\begin{aligned} \mathrm{E}X &= \mathrm{E}\left(\mathrm{E}(X|Y)\right) \\ &= \mathrm{E}(pY) && \text{(as before)} \\ &= \mathrm{E}\left(\mathrm{E}(pY|\Lambda)\right) \\ &= \mathrm{E}(p\Lambda) \\ &= p\beta, && \text{(exponential expectation)} \end{aligned}$$

completing the calculation. ‖

In this example we have used a slightly different type of model than before in that two of the random variables are discrete and one is continuous. Using these models should present no problems. We can define a joint density, $f(x, y, \lambda)$; conditional densities, $f(x|y)$, $f(x|y, \lambda)$, etc.; and marginal densities, $f(x)$, $f(x, y)$, etc. as before. Simply understand that, when probabilities or expectations are calculated, discrete variables are summed and continuous variables are integrated.

Note that this three-stage model can also be thought of as a two-stage hierarchy by combining the last two stages. If $Y|\Lambda \sim \text{Poisson}(\Lambda)$ and $\Lambda \sim \text{exponential}(\beta)$, then

$$\begin{aligned} P(Y=y) &= P(Y=y, 0<\Lambda<\infty) \\ &= \int_0^\infty f(y,\lambda)\,d\lambda \\ &= \int_0^\infty f(y|\lambda) f(\lambda)\,d\lambda \\ &= \int_0^\infty \left[\frac{e^{-\lambda}\lambda^y}{y!}\right]\frac{1}{\beta}e^{-\lambda/\beta}\,d\lambda \\ &= \frac{1}{\beta y!}\int_0^\infty \lambda^y e^{-\lambda(1+\beta^{-1})}\,d\lambda \qquad \begin{pmatrix}\text{gamma}\\ \text{pdf kernel}\end{pmatrix} \\ &= \frac{1}{\beta y!}\Gamma(y+1)\left(\frac{1}{1+\beta^{-1}}\right)^{y+1} \\ &= \frac{1}{(1+\beta)}\left(\frac{1}{1+\beta^{-1}}\right)^y. \end{aligned}$$

This expression for the pmf of Y is the form (3.2.10) of the negative binomial pmf. Therefore, our three-stage hierarchy in Example 4.4.5 is equivalent to the two-stage hierarchy

$$\begin{aligned} X|Y &\sim \text{binomial}(Y,p), \\ Y &\sim \text{negative binomial}\left(p=\frac{1}{1+\beta}, r=1\right). \end{aligned}$$

However, in terms of understanding the model, the three-stage model is much easier to understand!

A useful generalization is a Poisson–gamma mixture, which is a generalization of a part of the previous model. If we have the hierarchy

$$\begin{aligned} Y|\Lambda &\sim \text{Poisson}(\Lambda), \\ \Lambda &\sim \text{gamma}(\alpha,\beta), \end{aligned}$$

then the marginal distribution of Y is negative binomial (see Exercise 4.32). This model for the negative binomial distribution shows that it can be considered to be a "more variable" Poisson. Solomon (1983) explains these and other biological and mathematical models that lead to the negative binomial distribution. (See Exercise 4.33.)

Aside from the advantage in aiding understanding, hierarchical models can often make calculations easier. For example, a distribution that often occurs in statistics is the *noncentral chi squared distribution*. With p degrees of freedom and noncentrality parameter λ, the pdf is given by

$$f(x|\lambda,p) = \sum_{k=0}^\infty \frac{x^{p/2+k-1}e^{-x/2}}{\Gamma(p/2+k)2^{p/2+k}}\frac{\lambda^k e^{-\lambda}}{k!}, \tag{4.4.3}$$

an extremely messy expression. Calculating EX, for example, looks like quite a chore. However, if we examine the pdf closely, we see that this is a mixture distribution, made up of central chi squared densities (like those given in (3.2.10)) and Poisson distributions. That is, if we set up the hierarchy

$$X|K \sim \chi^2_{p+2K},$$
$$K \sim \text{Poisson}(\lambda),$$

then the marginal distribution of X is given by (4.4.3). Hence

$$\begin{aligned} EX &= E(E(X|K)) \\ &= E(p+2K) \\ &= p+2\lambda, \end{aligned}$$

a relatively simple calculation. $\text{Var}\, X$ can also be calculated in this way.

We close this section with one more hierarchical model and illustrate one more conditional expectation calculation.

Example 4.4.6 (Beta-binomial hierarchy) One generalization of the binomial distribution is to allow the success probability to vary according to a distribution. A standard model for this situation is

$$X|P \sim \text{binomial}(P), \quad i = 1, \ldots, n,$$
$$P \sim \text{beta}(\alpha, \beta).$$

By iterating the expectation, we calculate the mean of X as

$$EX = E[E(X|P)] = E[nP] = n\frac{\alpha}{\alpha+\beta}. \qquad \|$$

Calculating the variance of X is only slightly more involved. We can make use of a formula for conditional variances, similar in spirit to the expected value identity of Theorem 4.4.3.

Theorem 4.4.7 (Conditional variance identity) *For any two random variables X and Y,*

$$\text{Var}\, X = E\left(\text{Var}(X|Y)\right) + \text{Var}\left(E(X|Y)\right), \tag{4.4.4}$$

provided that the expectations exist.

Proof: By definition, we have

$$\text{Var}\, X = E\left([X - EX]^2\right) = E\left([X - E(X|Y) + E(X|Y) - EX]^2\right),$$

where in the last step we have added and subtracted $E(X|Y)$. Expanding the square in this last expectation now gives

$$\begin{aligned} \text{Var}\, X = {} & E\left([X - E(X|Y)]^2\right) + E\left([E(X|Y) - EX]^2\right) \\ & + 2E\left([X - E(X|Y)][E(X|Y) - EX]\right). \end{aligned} \tag{4.4.5}$$

The last term in this expression is equal to 0, however, which can easily be seen by iterating the expectation:

(4.4.6)

$$\mathrm{E}\left([X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]\right)=\mathrm{E}\left(\mathrm{E}\left\{[X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]|Y\right\}\right).$$

In the conditional distribution $X|Y$, X is the random variable. So in the expression

$$\mathrm{E}\left\{[X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]|Y\right\},$$

$\mathrm{E}(X|Y)$ and $\mathrm{E}X$ are constants. Thus,

$$\begin{aligned}\mathrm{E}\left\{[X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]|Y\right\}&=(\mathrm{E}(X|Y)-\mathrm{E}X)\left(\mathrm{E}\left\{[X-\mathrm{E}(X|Y)]|Y\right\}\right)\\&=(\mathrm{E}(X|Y)-\mathrm{E}X)\left(\mathrm{E}(X|Y)-\mathrm{E}(X|Y)\right)\\&=(\mathrm{E}(X|Y)-\mathrm{E}X)\,(0)\\&=0.\end{aligned}$$

Thus, from (4.4.6), we have that $\mathrm{E}((X-\mathrm{E}(X|Y))(\mathrm{E}(X|Y)-\mathrm{E}X))=\mathrm{E}(0)=0$. Referring back to equation (4.4.5), we see that

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

and

$$\mathrm{E}\left([\mathrm{E}(X|Y)-\mathrm{E}X]^2\right)=\mathrm{Var}\left(\mathrm{E}(X|Y)\right),$$

establishing (4.4.4). □

Example 4.4.8 (Continuation of Example 4.4.6) To calculate the variance of X, we have from (4.4.4),

$$\mathrm{Var}\,X=\mathrm{Var}\left(\mathrm{E}(X|P)\right)+\mathrm{E}\left(\mathrm{Var}(X|P)\right).$$

Now $\mathrm{E}(X|P)=nP$, and since $P\sim\text{beta}(\alpha,\beta)$,

$$\mathrm{Var}\left(\mathrm{E}(X|P)\right)=\mathrm{Var}(nP)=n^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Also, since $X|P$ is binomial(n,P), $\mathrm{Var}(X|P)=nP(1-P)$. We then have

$$\mathrm{E}\left[\mathrm{Var}(X|P)\right]=n\mathrm{E}\left[P(1-P)\right]=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p(1-p)p^{\alpha-1}(1-p)^{\beta-1}dp.$$

Notice that the integrand is the kernel of another beta pdf (with parameters $\alpha+1$ and $\beta+1$) so

$$\mathrm{E}\left(\mathrm{Var}(X|P)\right)=n\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left[\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}\right]=n\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}.$$

Adding together the two pieces and simplifying, we get

$$\mathrm{Var}\,X=n\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

‖

4.5 Covariance and Correlation

In earlier sections, we have discussed the absence or presence of a relationship between two random variables, independence or nonindependence. But if there is a relationship, the relationship may be strong or weak. In this section we discuss two numerical measures of the strength of a relationship between two random variables, the covariance and correlation.

To illustrate what we mean by the strength of a relationship between two random variables, consider two different experiments. In the first, random variables X and Y are measured, where X is the weight of a sample of water and Y is the volume of the same sample of water. Clearly there is a strong relationship between X and Y. If (X, Y) pairs are measured on several samples and the observed data pairs are plotted, the data points should fall on a straight line because of the physical relationship between X and Y. This will not be exactly the case because of measurement errors, impurities in the water, etc. But with careful laboratory technique, the data points will fall very nearly on a straight line. Now consider another experiment in which X and Y are measured, where X is the body weight of a human and Y is the same human's height. Clearly there is also a relationship between X and Y here but the relationship is not nearly as strong. We would not expect a plot of (X, Y) pairs measured on different people to form a straight line, although we might expect to see an upward trend in the plot. The covariance and correlation are two measures that quantify this difference in the strength of a relationship between two random variables.

Throughout this section we will frequently be referring to the mean and variance of X and the mean and variance of Y. For these we will use the notation $\mathrm{E}X = \mu_X$, $\mathrm{E}Y = \mu_Y$, $\mathrm{Var}\, X = \sigma_X^2$, and $\mathrm{Var}\, Y = \sigma_Y^2$. We will assume throughout that $0 < \sigma_X^2 < \infty$ and $0 < \sigma_Y^2 < \infty$.

Definition 4.5.1 The *covariance of X and Y* is the number defined by

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left((X - \mu_X)(Y - \mu_Y)\right).$$

Definition 4.5.2 The *correlation of X and Y* is the number defined by

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The value ρ_{XY} is also called the *correlation coefficient.*

If large values of X tend to be observed with large values of Y and small values of X with small values of Y, then $\mathrm{Cov}(X, Y)$ will be positive. If $X > \mu_X$, then $Y > \mu_Y$ is likely to be true and the product $(X - \mu_X)(Y - \mu_Y)$ will be positive. If $X < \mu_X$, then $Y < \mu_Y$ is likely to be true and the product $(X - \mu_X)(Y - \mu_Y)$ will again be positive. Thus $\mathrm{Cov}(X, Y) = \mathrm{E}(X - \mu_X)(Y - \mu_Y) > 0$. If large values of X tend to be observed with small values of Y and small values of X with large values of Y, then $\mathrm{Cov}(X, Y)$ will be negative because when $X > \mu_X$, Y will tend to be less than μ_Y and vice versa, and hence $(X - \mu_X)(Y - \mu_Y)$ will tend to be negative. Thus the sign of $\mathrm{Cov}(X, Y)$ gives information regarding the relationship between X and Y.

But $\text{Cov}(X,Y)$ can be any number and a given value of $\text{Cov}(X,Y)$, say $\text{Cov}(X,Y) = 3$, does not in itself give information about the strength of the relationship between X and Y. On the other hand, the correlation is always between -1 and 1, with the values -1 and 1 indicating a perfect *linear* relationship between X and Y. This is proved in Theorem 4.5.7.

Before investigating these properties of covariance and correlation, we will first calculate these measures in a given example. This calculation will be simplified by the following result.

Theorem 4.5.3 *For any random variables X and Y,*

$$\text{Cov}(X,Y) = \text{E}XY - \mu_X\mu_Y.$$

Proof:
$$\begin{aligned}
\text{Cov}(X,Y) &= \text{E}\,(X-\mu_X)(Y-\mu_Y) \\
&= \text{E}\,(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y) && \text{(expanding the product)} \\
&= \text{E}XY - \mu_X\text{E}Y - \mu_Y\text{E}X + \mu_X\mu_Y && (\mu_X \text{ and } \mu_Y \text{ are constants}) \\
&= \text{E}XY - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\
&= \text{E}XY - \mu_X\mu_Y.
\end{aligned}$$

□

Example 4.5.4 (Correlation–I) Let the joint pdf of (X,Y) be $f(x,y) = 1$, $0 < x < 1, x < y < x+1$. See Figure 4.5.1. The marginal distribution of X is uniform$(0,1)$ so $\mu_X = \frac{1}{2}$ and $\sigma_X^2 = \frac{1}{12}$. The marginal pdf of Y is $f_Y(y) = y$, $0 < y < 1$, and $f_Y(y) = 2-y, 1 \le y < 2$, with $\mu_Y = 1$ and $\sigma_Y^2 = \frac{1}{6}$. We also have

$$\begin{aligned}
\text{E}XY &= \int_0^1 \int_x^{x+1} xy\,dy\,dx = \int_0^1 \frac{1}{2}xy^2\Big|_x^{x+1} dx \\
&= \int_0^1 \left(x^2 + \frac{1}{2}x\right) dx = \frac{7}{12}.
\end{aligned}$$

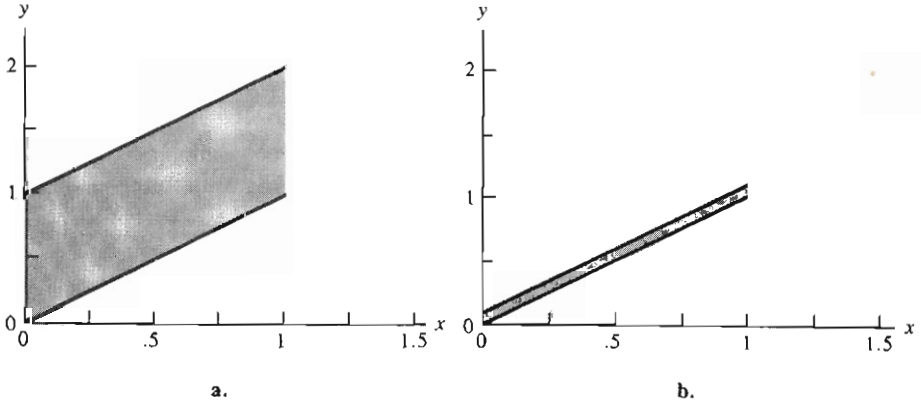

Figure 4.5.1. (a) *Region where $f(x,y) > 0$ for Example 4.5.4;* (b) *region where $f(x,y) > 0$ for Example 4.5.8*

Using Theorem 4.5.3, we have $\text{Cov}(X,Y) = \frac{7}{12} - \left(\frac{1}{2}\right)(1) = \frac{1}{12}$. The correlation is

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{1/12}{\sqrt{1/12}\sqrt{1/6}} = \frac{1}{\sqrt{2}}. \qquad \|$$

In the next three theorems we describe some of the fundamental properties of covariance and correlation.

Theorem 4.5.5 *If X and Y are independent random variables, then* $\text{Cov}(X,Y) = 0$ *and* $\rho_{XY} = 0$.

Proof: Since X and Y are independent, from Theorem 4.2.10 we have $EXY = (EX)(EY)$. Thus

$$\text{Cov}(X,Y) = EXY - (EX)(EY) = (EX)(EY) - (EX)(EY) = 0$$

and

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{0}{\sigma_X \sigma_Y} = 0. \qquad \square$$

Thus, the values $\text{Cov}(X,Y) = \rho_{XY} = 0$ in some sense indicate that there is no relationship between X and Y. It is important to note, however, that Theorem 4.5.5 does *not* say that if $\text{Cov}(X,Y) = 0$, then X and Y are independent. For example, if $X \sim f(x-\theta)$, symmetric around 0 with $EX = \theta$, and Y is the indicator function $Y = I(|X-\theta| < 2)$, then X and Y are obviously not independent. However,

$$E(XY) = \int_{-\infty}^{\infty} xI(|x-\theta| < 2) f(x-\theta)\, dx = \int_{-2}^{2} (t+\theta) f(t)\, dt = \theta \int_{-2}^{2} f(t)\, dt = EXEY,$$

where we used the fact that, by symmetry, $\int_{-2}^{2} t f(t)\, dt = 0$. So it is easy to find uncorrelated, dependent random variables.

Covariance and correlation measure only a particular kind of *linear* relationship that will be described further in Theorem 4.5.7. Also see Example 4.5.9, which discusses two random variables that have a strong relationship but whose covariance and correlation are 0 because the relationship is not linear.

Covariance also plays an important role in understanding the variation in sums of random variables, as the next theorem, a generalization of Theorem 2.3.4, indicates. (See Exercise 4.44 for a further generalization.)

Theorem 4.5.6 *If X and Y are any two random variables and a and b are any two constants, then*

$$\text{Var}(aX + bY) = a^2 \text{Var}\, X + b^2 \text{Var}\, Y + 2ab\, \text{Cov}(X,Y).$$

If X and Y are independent random variables, then

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

Proof: The mean of $aX + bY$ is $\mathrm{E}(aX + bY) = a\mathrm{E}X + b\mathrm{E}Y = a\mu_X + b\mu_Y$. Thus

$$\begin{aligned}\mathrm{Var}(aX + bY) &= \mathrm{E}\left((aX + bY) - (a\mu_X + b\mu_Y)\right)^2 \\ &= \mathrm{E}\left(a(X - \mu_X) + b(Y - \mu_Y)\right)^2 \\ &= \mathrm{E}\left(a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)\right) \\ &= a^2\mathrm{E}(X - \mu_X)^2 + b^2\mathrm{E}(Y - \mu_Y)^2 + 2ab\mathrm{E}(X - \mu_X)(Y - \mu_Y) \\ &= a^2\mathrm{Var}\,X + b^2\mathrm{Var}\,Y + 2ab\,\mathrm{Cov}(X, Y).\end{aligned}$$

If X and Y are independent, then, from Theorem 4.5.5, $\mathrm{Cov}(X, Y) = 0$ and the second equality is immediate from the first. □

From Theorem 4.5.6 we see that if X and Y are positively correlated ($\mathrm{Cov}(X, Y) > 0$), then the variation in $X + Y$ is greater than the sum of the variations in X and Y. But if they are negatively correlated, then the variation in $X + Y$ is less than the sum. For negatively correlated random variables, large values of one tend to be observed with small values of the other and in the sum these two extremes cancel. The result, $X + Y$, tends not to have as many extreme values and hence has smaller variance. By choosing $a = 1$ and $b = -1$ we get an expression for the variance of the difference of two random variables, and similar arguments apply.

The nature of the linear relationship measured by covariance and correlation is somewhat explained by the following theorem.

Theorem 4.5.7 *For any random variables X and Y,*

a. $-1 \le \rho_{XY} \le 1$.

b. $|\rho_{XY}| = 1$ *if and only if there exist numbers $a \neq 0$ and b such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$.*

Proof: Consider the function $h(t)$ defined by

$$h(t) = \mathrm{E}\left((X - \mu_X)t + (Y - \mu_Y)\right)^2.$$

Expanding this expression, we obtain

$$\begin{aligned}h(t) &= t^2\mathrm{E}(X - \mu_X)^2 + 2t\mathrm{E}(X - \mu_X)(Y - \mu_Y) + \mathrm{E}(Y - \mu_Y)^2 \\ &= t^2\sigma_X^2 + 2t\mathrm{Cov}(X, Y) + \sigma_Y^2.\end{aligned}$$

This quadratic function of t is greater than or equal to 0 for all values of t since it is the expected value of a nonnegative random variable. Thus, this quadratic function can have at most one real root and thus must have a nonpositive discriminant. That is,

$$(2\mathrm{Cov}(X, Y))^2 - 4\sigma_X^2\sigma_Y^2 \le 0.$$

This is equivalent to

$$-\sigma_X\sigma_Y \le \mathrm{Cov}(X, Y) \le \sigma_X\sigma_Y.$$

Dividing by $\sigma_X\sigma_Y$ yields

$$-1 \le \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \rho_{XY} \le 1.$$

Also, $|\rho_{XY}| = 1$ if and only if the discriminant is equal to 0. That is, $|\rho_{XY}| = 1$ if and only if $h(t)$ has a single root. But since $((X-\mu_X)t + (Y-\mu_Y))^2 \ge 0$, the expected value $h(t) = \text{E}((X-\mu_X)t + (Y-\mu_Y))^2 = 0$ if and only if

$$P\big([(X-\mu_X)t + (Y-\mu_Y)]^2 = 0\big) = 1.$$

This is equivalent to

$$P((X-\mu_X)t + (Y-\mu_Y) = 0) = 1.$$

This is $P(Y = aX + b) = 1$ with $a = -t$ and $b = \mu_X t + \mu_Y$, where t is the root of $h(t)$. Using the quadratic formula, we see that this root is $t = -\text{Cov}(X,Y)/\sigma_X^2$. Thus $a = -t$ has the same sign as ρ_{XY}, proving the final assertion. □

In Section 4.7 we will prove a theorem called the Cauchy–Schwarz Inequality. This theorem has as a direct consequence that ρ_{XY} is bounded between -1 and 1, and we will see that, with this inequality, the preceding proof can be shortened.

If there is a line $y = ax + b$, with $a \neq 0$, such that the values of (X,Y) have a high probability of being near this line, then the correlation between X and Y will be near 1 or -1. But if no such line exists, the correlation will be near 0. This is an intuitive notion of the linear relationship that is being measured by correlation. This idea will be illustrated further in the next two examples.

Example 4.5.8 (Correlation–II) This example is similar to Example 4.5.4, but we develop it differently to illustrate other model building and computational techniques. Let X have a uniform$(0,1)$ distribution and Z have a uniform$(0, \frac{1}{10})$ distribution. Suppose X and Z are independent. Let $Y = X + Z$ and consider the random vector (X,Y). The joint distribution of (X,Y) can be derived from the joint distribution of (X,Z) using the techniques of Section 4.3. The joint pdf of (X,Y) is

$$f(x,y) = 10, \quad 0 < x < 1, \quad x < y < x + \frac{1}{10}.$$

Rather than using the formal techniques of Section 4.3, we can justify this as follows. Given $X = x$, $Y = x + Z$. The conditional distribution of Z given $X = x$ is just uniform$(0, \frac{1}{10})$ since X and Z are independent. Thus x serves as a location parameter in the conditional distribution of Y given $X = x$, and this conditional distribution is just uniform$(x, x + \frac{1}{10})$. Multiplying this conditional pdf by the marginal pdf of X (uniform$(0,1)$) yields the joint pdf above. This representation of $Y = X + Z$ makes the computation of the covariance and correlation easy. The expected values of X and Y are $\text{E}X = \frac{1}{2}$ and $\text{E}Y = \text{E}(X+Z) = \text{E}X + \text{E}Z = \frac{1}{2} + \frac{1}{20} = \frac{11}{20}$, giving

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

$$= EX^2 + EXZ - (EX)^2 - (EX)(EZ)$$

$$= EX^2 - (EX)^2 + (EX)(EZ) - (EX)(EZ) \quad \begin{pmatrix}\text{independence of} \\ X \text{ and } Z\end{pmatrix}$$

$$= \sigma_X^2 = \frac{1}{12}.$$

From Theorem 4.5.6, the variance of Y is $\sigma_Y^2 = \text{Var}(X + Z) = \text{Var}\, X + \text{Var}\, Z = \frac{1}{12} + \frac{1}{1200}$. Thus

$$\rho_{XY} = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}}\sqrt{\frac{1}{12} + \frac{1}{1200}}} = \sqrt{\frac{100}{101}}.$$

This is much larger than the value of $\rho_{XY} = 1/\sqrt{2}$ obtained in Example 4.5.4. The sets on which $f(x, y)$ is positive for Example 4.5.4 and this example are illustrated in Figure 4.5.1. (Recall that this set is called the support of a distribution.) In each case, (X, Y) is a random point from the set. In both cases there is a linearly increasing relationship between X and Y, but the relationship is much stronger in Figure 4.5.1b. Another way to see this is by noting that in this example, the conditional distribution of Y given $X = x$ is uniform$(x, x + \frac{1}{10})$. In Example 4.5.4, the conditional distribution of Y given $X = x$ is uniform$(x, x + 1)$. The knowledge that $X = x$ gives us much more information about the value of Y in this model than in the one in Example 4.5.4. Hence the correlation is nearer to 1 in this example. ‖

The next example illustrates that there may be a strong relationship between X and Y, but if the relationship is not linear, the correlation may be small.

Example 4.5.9 (Correlation–III) In this example, let X have a uniform$(-1, 1)$ distribution and let Z have a uniform$(0, \frac{1}{10})$ distribution. Let X and Z be independent. Let $Y = X^2 + Z$ and consider the random vector (X, Y). As in Example 4.5.8, given $X = x$, $Y = x^2 + Z$ and the conditional distribution of Y given $X = x$ is uniform$(x^2, x^2 + \frac{1}{10})$. The joint pdf of X and Y, the product of this conditional pdf and the marginal pdf of X, is thus

$$f(x, y) = 5, \quad -1 < x < 1, \quad x^2 < y < x^2 + \frac{1}{10}.$$

The set on which $f(x, y) > 0$ is illustrated in Figure 4.5.2. There is a strong relationship between X and Y, as indicated by the conditional distribution of Y given $X = x$. But the relationship is not linear. The possible values of (X, Y) cluster around a parabola rather than a straight line. The correlation does not measure this nonlinear relationship. In fact, $\rho_{XY} = 0$. Since X has a uniform$(-1, 1)$ distribution, $EX = EX^3 = 0$, and since X and Z are independent, $EXZ = (EX)(EZ)$. Thus,

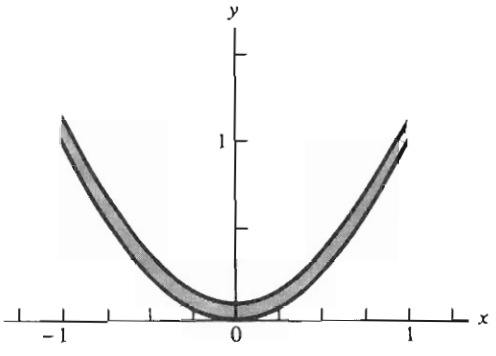

Figure 4.5.2. *Region where* $f(x, y) > 0$ *for Example 4.5.9*

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathrm{E}(X(X^2 + Z)) - (\mathrm{E}X)(\mathrm{E}(X^2 + Z)) \\ &= \mathrm{E}X^3 + \mathrm{E}XZ - 0\mathrm{E}(X^2 + Z) \\ &= 0 + (\mathrm{E}X)(\mathrm{E}Z) = 0(\mathrm{E}Z) = 0,\end{aligned}$$

and $\rho_{XY} = \mathrm{Cov}(X, Y)/(\sigma_X \sigma_Y) = 0$. ‖

We close this section by introducing a very important bivariate distribution in which the correlation coefficient arises naturally as a parameter.

Definition 4.5.10 Let $-\infty < \mu_X < \infty, -\infty < \mu_Y < \infty$, $0 < \sigma_X$, $0 < \sigma_Y$, and $-1 < \rho < 1$ be five real numbers. The *bivariate normal pdf with means* μ_X *and* μ_Y, *variances* σ_X^2 *and* σ_Y^2, *and correlation* ρ is the bivariate pdf given by

$$\begin{aligned}f(x, y) = &\left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \\ &\times \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right.\right. \\ &\left.\left. - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)\end{aligned}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

Although the formula for the bivariate normal pdf looks formidable, this bivariate distribution is one of the most frequently used. (In fact, the derivation of the formula need not be formidable at all. See Exercise 4.46.)

The many nice properties of this distribution include these:

a. The marginal distribution of X is $\mathrm{n}(\mu_X, \sigma_X^2)$.

b. The marginal distribution of Y is $\mathrm{n}(\mu_Y, \sigma_Y^2)$.

c. The correlation between X and Y is $\rho_{XY} = \rho$.

d. For any constants a and b, the distribution of $aX + bY$ is $\mathrm{n}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$.

We will leave the verification of properties (a), (b), and (d) as exercises (Exercise 4.45). Assuming (a) and (b) are true, we will prove (c). We have by definition

$$\begin{aligned}\rho_{XY} &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} \\ &= \frac{\mathrm{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y} \\ &= \mathrm{E}\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) f(x,y)\,dx\,dy.\end{aligned}$$

Make the change of variable

$$s = \left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) \quad \text{and} \quad t = \left(\frac{x-\mu_X}{\sigma_X}\right).$$

Then $x = \sigma_X t + \mu_X$, $y = (\sigma_Y s/t) + \mu_Y$, and the Jacobian of the transformation is $J = \sigma_X\sigma_Y/t$. With this change of variable, we obtain

$$\begin{aligned}\rho_{XY} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} sf\left(\sigma_X t + \mu_X, \frac{\sigma_Y s}{t} + \mu_Y\right)\left|\frac{\sigma_X\sigma_Y}{t}\right| ds\,dt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s\left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1} \\ &\qquad \times \exp\left(-\frac{1}{2(1-\rho^2)}\left(t^2 - 2\rho s + \left(\frac{s}{t}\right)^2\right)\right)\frac{\sigma_X\sigma_Y}{|t|}\,ds\,dt.\end{aligned}$$

Noting that $|t| = \sqrt{t^2}$ and $t^2 - 2\rho s + \left(\frac{s}{t}\right)^2 = \left(\frac{s-\rho t^2}{t}\right)^2 + (1-\rho^2)t^2$, we can rewrite this as

$$\rho_{XY} = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)\left[\int_{-\infty}^{\infty}\frac{s}{\sqrt{2\pi}\sqrt{(1-\rho^2)t^2}}\exp\left(-\frac{(s-\rho t^2)^2}{2(1-\rho^2)t^2}\right)ds\right]dt.$$

The inner integral is $\mathrm{E}S$, where S is a normal random variable with $\mathrm{E}S = \rho t^2$ and $\mathrm{Var}\,S = (1-\rho^2)t^2$. Thus the inner integral is ρt^2. Hence we have

$$\rho_{XY} = \int_{-\infty}^{\infty}\frac{\rho t^2}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)dt.$$

But this integral is $\rho\mathrm{E}T^2$, where T is a $\mathrm{n}(0,1)$ random variable. Hence $\mathrm{E}T^2 = 1$ and $\rho_{XY} = \rho$.

All the conditional distributions of Y given $X = x$ and of X given $Y = y$ are also normal distributions. Using the joint and marginal pdfs given above, it is straightforward to verify that the conditional distribution of Y given $X = x$ is

$$\mathrm{n}(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \sigma_Y^2(1 - \rho^2)).$$

As ρ converges to 1 or -1, the conditional variance $\sigma_Y^2(1-\rho^2)$ converges to 0. Thus, the conditional distribution of Y given $X = x$ becomes more concentrated about the point $\mu_Y+\rho(\sigma_Y/\sigma_X)(x-\mu_X)$, and the joint probability distribution of (X, Y) becomes more concentrated about the line $y = \mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X)$. This illustrates again the point made earlier that a correlation near 1 or -1 means that there is a line $y = ax + b$ about which the values of (X, Y) cluster with high probability.

Note one important fact: All of the normal marginal and conditional pdfs are derived from the starting point of bivariate normality. The derivation does not go in the opposite direction. That is, marginal normality does not imply joint normality. See Exercise 4.47 for an illustration of this.

4.6 Multivariate Distributions

At the beginning of this chapter, we discussed observing more than two random variables in an experiment. In the previous sections our discussions have concentrated on a bivariate random vector (X, Y). In this section we discuss a multivariate random vector $(X_1, \ldots, X_n)$. In the example at the beginning of this chapter, temperature, height, weight, and blood pressure were observed on an individual. In this example, $n = 4$ and the observed random vector is (X_1, X_2, X_3, X_4), where X_1 is temperature, X_2 is height, etc. The concepts from the earlier sections, including marginal and conditional distributions, generalize from the bivariate to the multivariate setting. We introduce some of these generalizations in this section.

A note on notation: We will use boldface letters to denote multiple variates. Thus, we write $\mathbf{X}$ to denote the random variables $X_1, \ldots, X_n$ and $\mathbf{x}$ to denote the sample $x_1, \ldots, x_n$.

The random vector $\mathbf{X} = (X_1, \ldots, X_n)$ has a sample space that is a subset of $\Re^n$. If $(X_1, \ldots, X_n)$ is a discrete random vector (the sample space is countable), then the *joint pmf of* $(X_1, \ldots, X_n)$ is the function defined by $f(\mathbf{x}) = f(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$ for each $(x_1, \ldots, x_n) \in \Re^n$. Then for any $A \subset \Re^n$,

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} f(\mathbf{x}). \tag{4.6.1}$$

If $(X_1, \ldots, X_n)$ is a continuous random vector, the *joint pdf of* $(X_1, \ldots, X_n)$ is a function $f(x_1, \ldots, x_n)$ that satisfies

$$P(\mathbf{X} \in A) = \int \cdots \int_A f(\mathbf{x}) d\mathbf{x} = \int \cdots \int_A f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n. \tag{4.6.2}$$

These integrals are n-fold integrals with limits of integration set so that the integration is over all points $\mathbf{x} \in A$.

Let $g(\mathbf{x}) = g(x_1, \ldots, x_n)$ be a real-valued function defined on the sample space of $\mathbf{X}$. Then $g(\mathbf{X})$ is a random variable and the *expected value of* $g(\mathbf{X})$ is

$$(4.6.3) \quad \mathrm{E}g(\mathbf{X}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad \text{and} \quad \mathrm{E}g(\mathbf{X}) = \sum_{\mathbf{x} \in \Re^n} g(\mathbf{x}) f(\mathbf{x})$$

in the continuous and discrete cases, respectively. These and other definitions are analogous to the bivariate definitions except that now the integrals or sums are over the appropriate subset of $\Re^n$ rather than $\Re^2$.

The *marginal pdf or pmf* of any subset of the coordinates of $(X_1, \ldots, X_n)$ can be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates. Thus, for example, the marginal distribution of $(X_1, \ldots, X_k)$, the first k coordinates of $(X_1, \ldots, X_n)$, is given by the pdf or pmf

$$(4.6.4) \qquad f(x_1, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_{k+1} \cdots dx_n$$

or

$$(4.6.5) \qquad f(x_1, \ldots, x_k) = \sum_{(x_{k+1}, \ldots, x_n) \in \Re^{n-k}} f(x_1, \ldots, x_n)$$

for every $(x_1, \ldots, x_k) \in \Re^k$. The *conditional pdf or pmf* of a subset of the coordinates of $(X_1, \ldots, X_n)$ given the values of the remaining coordinates is obtained by dividing the joint pdf or pmf by the marginal pdf or pmf of the remaining coordinates. Thus, for example, if $f(x_1, \ldots, x_k) > 0$, the conditional pdf or pmf of $(X_{k+1}, \ldots, X_n)$ given $X_1 = x_1, \ldots, X_k = x_k$ is the function of $(x_{k+1}, \ldots, x_n)$ defined by

$$(4.6.6) \qquad f(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_k)}.$$

These ideas are illustrated in the following example.

Example 4.6.1 (Multivariate pdfs) Let $n = 4$ and

$$f(x_1, x_2, x_3, x_4) = \begin{cases} \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2) & 0 < x_i < 1, i = 1, 2, 3, 4 \\ 0 & \text{otherwise.} \end{cases}$$

This nonnegative function is the joint pdf of a random vector (X_1, X_2, X_3, X_4) and it can be verified that

$$\begin{aligned} &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_1\, dx_2\, dx_3\, dx_4 \\ &\quad = \int_0^1 \int_0^1 \int_0^1 \int_0^1 \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2)\, dx_1\, dx_2\, dx_3\, dx_4 \\ &\quad = 1. \end{aligned}$$

This joint pdf can be used to compute probabilities such as

$$P\left(X_1 < \frac{1}{2}, X_2 < \frac{3}{4}, X_4 > \frac{1}{2}\right)$$
$$= \int_{\frac{1}{2}}^{1}\int_{0}^{1}\int_{0}^{\frac{3}{4}}\int_{0}^{\frac{1}{2}} \frac{3}{4}\left(x_1^2 + x_2^2 + x_3^2 + x_4^2\right) dx_1\, dx_2\, dx_3\, dx_4.$$

Note how the limits of integration restrict the integration to those values of (x_1, x_2, x_3, x_4) that are in the event in question and for which $f(x_1, x_2, x_3, x_4) > 0$. Each of the four terms, $\frac{3}{4}x_1^2$, $\frac{3}{4}x_2^2$, etc., can be integrated separately and the results summed. For example,

$$\int_{\frac{1}{2}}^{1}\int_{0}^{1}\int_{0}^{\frac{3}{4}}\int_{0}^{\frac{1}{2}} \frac{3}{4}x_1^2\, dx_1\, dx_2\, dx_3\, dx_4 = \frac{3}{256}.$$

The other three integrals are $\frac{7}{1024}$, $\frac{3}{64}$, and $\frac{21}{256}$. Thus

$$P\left(X_1 < \frac{1}{2}, X_2 < \frac{3}{4}, X_4 > \frac{1}{2}\right) = \frac{3}{256} + \frac{7}{1024} + \frac{3}{64} + \frac{21}{256} = \frac{151}{1024}.$$

Using (4.6.4), we can obtain the marginal pdf of (X_1, X_2) by integrating out the variables x_3 and x_4 to obtain

$$f(x_1, x_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4)\, dx_3\, dx_4$$
$$= \int_{0}^{1}\int_{0}^{1} \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2)\, dx_3\, dx_4 = \frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2}$$

for $0 < x_1 < 1$ and $0 < x_2 < 1$. Any probability or expected value that involves only X_1 and X_2 can be computed using this marginal pdf. For example,

$$\mathrm{E}X_1X_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f(x_1, x_2)\, dx_1\, dx_2$$
$$= \int_{0}^{1}\int_{0}^{1} x_1x_2\left(\frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2}\right) dx_1\, dx_2$$
$$= \int_{0}^{1}\int_{0}^{1}\left(\frac{3}{4}x_1^3x_2 + \frac{3}{4}x_1x_2^3 + \frac{1}{2}x_1x_2\right) dx_1\, dx_2$$
$$= \int_{0}^{1}\left(\frac{3}{16}x_2 + \frac{3}{8}x_2^3 + \frac{1}{4}x_2\right) dx_2 = \frac{3}{32} + \frac{3}{32} + \frac{1}{8} = \frac{5}{16}.$$

For any (x_1, x_2) with $0 < x_1 < 1$ and $0 < x_2 < 1$, $f(x_1, x_2) > 0$ and the conditional pdf of (X_3, X_4) given $X_1 = x_1$ and $X_2 = x_2$ can be found using (4.6.6). For any such (x_1, x_2), $f(x_1, x_2, x_3, x_4) > 0$ if $0 < x_3 < 1$ and $0 < x_4 < 1$, and for these values of (x_3, x_4), the conditional pdf is

$$f(x_3, x_4|x_1, x_2) = \frac{f(x_1, x_2, x_3, x_4)}{f(x_1, x_2)}$$
$$= \frac{\frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2)}{\frac{3}{4}(x_1^2 + x_2^2) + \frac{1}{2}}$$
$$= \frac{x_1^2 + x_2^2 + x_3^2 + x_4^2}{x_1^2 + x_2^2 + \frac{2}{3}}.$$

For example, the conditional pdf of (X_3, X_4) given $X_1 = \frac{1}{3}$ and $X_2 = \frac{2}{3}$ is

$$f\left(x_3, x_4 \middle| x_1 = \frac{1}{3}, x_2 = \frac{2}{3}\right) = \frac{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + x_3^2 + x_4^2}{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \frac{2}{3}} = \frac{5}{11} + \frac{9}{11}x_3^2 + \frac{9}{11}x_4^2.$$

This can be used to compute

$$P\left(X_3 > \frac{3}{4}, X_4 < \frac{1}{2} \middle| X_1 = \frac{1}{3}, X_2 = \frac{2}{3}\right) = \int_0^{\frac{1}{2}} \int_{\frac{3}{4}}^1 \left(\frac{5}{11} + \frac{9}{11}x_3^2 + \frac{9}{11}x_4^2\right) dx_3\, dx_4$$
$$= \int_0^{\frac{1}{2}} \left(\frac{5}{44} + \frac{111}{704} + \frac{9}{44}x_4^2\right) dx_4$$
$$= \frac{5}{88} + \frac{111}{1408} + \frac{3}{352} = \frac{203}{1408}. \quad \|$$

Before giving examples of computations with conditional and marginal distributions for a discrete multivariate random vector, we will introduce an important family of discrete multivariate distributions. This family generalizes the binomial family to the situation in which each trial has n (rather than two) distinct possible outcomes.

Definition 4.6.2 Let n and m be positive integers and let $p_1, \ldots, p_n$ be numbers satisfying $0 \le p_i \le 1$, $i = 1, \ldots, n$, and $\sum_{i=1}^n p_i = 1$. Then the random vector $(X_1, \ldots, X_n)$ has a *multinomial distribution with m trials and cell probabilities* $p_1, \ldots, p_n$ if the joint pmf of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n) = \frac{m!}{x_1! \cdot \cdots \cdot x_n!} p_1^{x_1} \cdot \cdots \cdot p_n^{x_n} = m! \prod_{i=1}^n \frac{p_i^{x_i}}{x_i!}$$

on the set of $(x_1, \ldots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$.

The multinomial distribution is a model for the following kind of experiment. The experiment consists of m independent trials. Each trial results in one of n distinct possible outcomes. The probability of the ith outcome is p_i on every trial. And X_i is the count of the number of times the ith outcome occurred in the m trials. For $n = 2$, this is just a binomial experiment in which each trial has $n = 2$ possible outcomes and X_1 counts the number of "successes" and $X_2 = m - X_1$ counts the number of "failures" in m trials. In a general multinomial experiment, there are n different possible outcomes to count.

Example 4.6.3 (Multivariate pmf) Consider tossing a six-sided die ten times. Suppose the die is unbalanced so that the probability of observing a 1 is $\frac{1}{21}$, the probability of observing a 2 is $\frac{2}{21}$, and, in general, the probability of observing an i is $\frac{i}{21}$. Now consider the random vector $(X_1, \ldots, X_6)$, where X_i counts the number of times i comes up in the ten tosses. Then $(X_1, \ldots, X_6)$ has a multinomial distribution with $m = 10$ trials, $n = 6$ possible outcomes, and cell probabilities $p_1 = \frac{1}{21}, p_2 = \frac{2}{21}, \ldots, p_6 = \frac{6}{21}$. The formula in Definition 4.6.2 may be used to calculate the probability of rolling four 6s, three 5s, two 4s, and one 3 to be

$$f(0,0,1,2,3,4) = \frac{10!}{0!0!1!2!3!4!}\left(\frac{1}{21}\right)^0\left(\frac{2}{21}\right)^0\left(\frac{3}{21}\right)^1\left(\frac{4}{21}\right)^2\left(\frac{5}{21}\right)^3\left(\frac{6}{21}\right)^4$$
$$= .0059. \qquad \|$$

The factor $m!/(x_1!\cdots x_n!)$ is called a *multinomial coefficient.* It is the number of ways that m objects can be divided into n groups with x_1 in the first group, x_2 in the second group, ..., and x_n in the nth group. A generalization of the Binomial Theorem 3.2.2 is the Multinomial Theorem.

Theorem 4.6.4 (Multinomial Theorem) *Let m and n be positive integers. Let $\mathcal{A}$ be the set of vectors $\mathbf{x} = (x_1, \ldots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^n x_i = m$. Then, for any real numbers $p_1, \ldots, p_n$,*

$$(p_1 + \cdots + p_n)^m = \sum_{\mathbf{x}\in\mathcal{A}} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n}.$$

Theorem 4.6.4 shows that a multinomial pmf sums to 1. The set $\mathcal{A}$ is the set of points with positive probability in Definition 4.6.2. The sum of the pmf over all those points is, by Theorem 4.6.4, $(p_1 + \cdots + p_n)^m = 1^m = 1$.

Now we consider some marginal and conditional distributions for the multinomial model. Consider a single coordinate X_i. If the occurrence of the ith outcome is labeled a "success" and anything else is labeled a "failure," then X_i is the count of the number of successes in m independent trials where the probability of a success is p_i on each trial. Thus X_i should have a binomial(m, p_i) distribution. To verify this the marginal distribution of X_i should be computed using (4.6.5). For example, consider the marginal pmf of X_n. For a fixed value of $x_n \in \{0, 1, \ldots, n\}$, to compute the marginal pmf $f(x_n)$, we must sum over all possible values of $(x_1, \ldots, x_{n-1})$. That is, we must sum over all $(x_1, \ldots, x_{n-1})$ such that the x_is are all nonnegative integers and $\sum_{i=1}^{n-1} x_i = m - x_n$. Denote this set by $\mathcal{B}$. Then

$$f(x_n) = \sum_{(x_1,\ldots,x_{n-1})\in\mathcal{B}} \frac{m!}{x_1!\cdots x_n!}(p_1)^{x_1}\cdots(p_n)^{x_n}$$
$$= \sum_{(x_1,\ldots,x_{n-1})\in\mathcal{B}} \frac{m!}{x_1!\cdots x_n!} p_1^{x_1}\cdots p_n^{x_n} \frac{(m-x_n)!}{(m-x_n)!}\frac{(1-p_n)^{m-x_n}}{(1-p_n)^{m-x_n}}$$

$$= \frac{m!}{x_n!(m-x_n)!} p_n^{x_n}(1-p_n)^{m-x_n}$$
$$\times \sum_{(x_1,\ldots,x_{n-1})\in\mathcal{B}} \frac{(m-x_n)!}{x_1!\cdots\cdot x_{n-1}!}\left(\frac{p_1}{1-p_n}\right)^{x_1}\cdots\cdot\left(\frac{p_{n-1}}{1-p_n}\right)^{x_{n-1}}.$$

But using the facts that $x_1+\cdots+x_{n-1}=m-x_n$ and $p_1+\cdots+p_{n-1}=1-p_n$ and Theorem 4.6.4, we see that the last summation is 1. Hence the marginal distribution of X_n is binomial(m,p_n). Similar arguments show that each of the other coordinates is marginally binomially distributed.

Given that $X_n=x_n$, there must have been $m-x_n$ trials that resulted in one of the first $n-1$ outcomes. The vector $(X_1,\ldots,X_{n-1})$ counts the number of these $m-x_n$ trials that are of each type. Thus it seems that given $X_n=x_n$, $(X_1,\ldots,X_{n-1})$ might have a multinomial distribution. This is true. From (4.6.6), the conditional pmf of $(X_1,\ldots,X_{n-1})$ given $X_n=x_n$ is

$$f(x_1,\ldots,x_{n-1}|x_n) = \frac{f(x_1,\ldots,x_n)}{f(x_n)}$$
$$= \frac{\frac{m!}{x_1!\cdots\cdot x_n!}(p_1)^{x_1}\cdots\cdot(p_n)^{x_n}}{\frac{m!}{x_n!(m-x_n)!}(p_n)^{x_n}(1-p_n)^{m-x_n}}$$
$$= \frac{(m-x_n)!}{x_1!\cdots\cdot x_{n-1}!}\left(\frac{p_1}{1-p_n}\right)^{x_1}\cdots\cdot\left(\frac{p_{n-1}}{1-p_n}\right)^{x_{n-1}}.$$

This is the pmf of a multinomial distribution with $m-x_n$ trials and cell probabilities $p_1/(1-p_n),\ldots,p_{n-1}/(1-p_n)$. In fact, the conditional distribution of any subset of the coordinates of $(X_1,\ldots,X_n)$ given the values of the rest of the coordinates is a multinomial distribution.

We see from the conditional distributions that the coordinates of the vector $(X_1,\ldots,X_n)$ are related. In particular, there must be some negative correlation. It turns out that all of the pairwise covariances are negative and are given by (Exercise 4.39)

$$\mathrm{Cov}(X_i,X_j)=\mathrm{E}[(X_i-p_i)(X_j-p_j)]=-mp_ip_j.$$

Thus, the negative correlation is greater for variables with higher success probabilities. This makes sense, as the variable total is constrained at m, so if one starts to get big, the other tends not to.

Definition 4.6.5 Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be random vectors with joint pdf or pmf $f(\mathbf{x}_1,\ldots,\mathbf{x}_n)$. Let $f_{\mathbf{X}_i}(\mathbf{x}_i)$ denote the marginal pdf or pmf of $\mathbf{X}_i$. Then $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are called *mutually independent random vectors* if, for every $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$,

$$f(\mathbf{x}_1,\ldots,\mathbf{x}_n)=f_{\mathbf{X}_1}(\mathbf{x}_1)\cdot\cdots\cdot f_{\mathbf{X}_n}(\mathbf{x}_n)=\prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i).$$

If the X_is are all one-dimensional, then $X_1,\ldots,X_n$ are called *mutually independent random variables.*

If $X_1, \ldots, X_n$ are mutually independent, then knowledge about the values of some coordinates gives us no information about the values of the other coordinates. Using Definition 4.6.5, one can show that the conditional distribution of any subset of the coordinates, given the values of the rest of the coordinates, is the same as the marginal distribution of the subset. Mutual independence implies that any pair, say X_i and X_j, are pairwise independent. That is, the bivariate marginal pdf or pmf, $f(x_i, x_j)$, satisfies Definition 4.2.5. But mutual independence implies more than pairwise independence. As in Example 1.3.11, it is possible to specify a probability distribution for $(X_1, \ldots, X_n)$ with the property that each pair, (X_i, X_j), is pairwise independent but $X_1, \ldots, X_n$ are not mutually independent.

Mutually independent random variables have many nice properties. The proofs of the following theorems are analogous to the proofs of their counterparts in Sections 4.2 and 4.3.

Theorem 4.6.6 (Generalization of Theorem 4.2.10) *Let $X_1, \ldots, X_n$ be mutually independent random variables. Let $g_1, \ldots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of x_i, $i = 1, \ldots, n$. Then*

$$\mathrm{E}(g_1(X_1)\cdot\cdots\cdot g_n(X_n)) = (\mathrm{E}g_1(X_1))\cdot\cdots\cdot(\mathrm{E}g_n(X_n)).$$

Theorem 4.6.7 (Generalization of Theorem 4.2.12) *Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $Z = X_1 + \cdots + X_n$. Then the mgf of Z is*

$$M_Z(t) = M_{X_1}(t)\cdot\cdots\cdot M_{X_n}(t).$$

In particular, if $X_1, \ldots, X_n$ all have the same distribution with mgf $M_X(t)$, then

$$M_Z(t) = (M_X(t))^n.$$

Example 4.6.8 (Mgf of a sum of gamma variables) Suppose $X_1, \ldots, X_n$ are mutually independent random variables, and the distribution of X_i is gamma(α_i, β). From Example 2.3.8, the mgf of a gamma(α, β) distribution is $M(t) = (1 - \beta t)^{-\alpha}$. Thus, if $Z = X_1 + \cdots + X_n$, the mgf of Z is

$$M_Z(t) = M_{X_1}(t)\cdot\cdots\cdot M_{X_n}(t) = (1-\beta t)^{-\alpha_1}\cdot\cdots\cdot(1-\beta t)^{-\alpha_n} = (1-\beta t)^{-(\alpha_1+\cdots+\alpha_n)}.$$

This is the mgf of a gamma$(\alpha_1 + \cdots + \alpha_n, \beta)$ distribution. Thus, the sum of independent gamma random variables that have a common scale parameter β also has a gamma distribution. ‖

A generalization of Theorem 4.6.7 is obtained if we consider a sum of linear functions of independent random variables.

Corollary 4.6.9 *Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Let $Z = (a_1X_1 + b_1) + \cdots + (a_nX_n + b_n)$. Then the mgf of Z is*

$$M_Z(t) = (e^{t(\Sigma b_i)})M_{X_1}(a_1t)\cdot\cdots\cdot M_{X_n}(a_nt).$$

Proof: From the definition, the mgf of Z is

$$\begin{aligned} M_Z(t) &= \mathrm{E}e^{tZ} \\ &= \mathrm{E}e^{t\Sigma(a_iX_i+b_i)} \\ &= (e^{t(\Sigma b_i)})\mathrm{E}(e^{ta_1X_1}\cdot\cdots\cdot e^{ta_nX_n}) \qquad \left(\begin{array}{c}\text{properties of exponentials}\\ \text{and expectations}\end{array}\right) \\ &= (e^{t(\Sigma b_i)})M_{X_1}(a_1t)\cdot\cdots\cdot M_{X_n}(a_nt), \qquad \text{(Theorem 4.6.6)} \end{aligned}$$

as was to be shown. □

Undoubtedly, the most important application of Corollary 4.6.9 is to the case of normal random variables. *A linear combination of independent normal random variables is normally distributed.*

Corollary 4.6.10 *Let $X_1, \ldots, X_n$ be mutually independent random variables with $X_i \sim \mathrm{n}(\mu_i, \sigma_i^2)$. Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Then*

$$Z = \sum_{i=1}^{n}(a_iX_i + b_i) \sim \mathrm{n}\left(\sum_{i=1}^{n}(a_i\mu_i + b_i), \sum_{i=1}^{n}a_i^2\sigma_i^2\right).$$

Proof: Recall that the mgf of a $\mathrm{n}(\mu, \sigma^2)$ random variable is $M(t) = e^{\mu t+\sigma^2t^2/2}$. Substituting into the expression in Corollary 4.6.9 yields

$$\begin{aligned} M_Z(t) &= (e^{t(\Sigma b_i)})e^{\mu_1a_1t+\sigma_1^2a_1^2t^2/2}\cdot\cdots\cdot e^{\mu_na_nt+\sigma_n^2a_n^2t^2/2} \\ &= e^{\left((\Sigma(a_i\mu_i+b_i))t+\left(\Sigma a_i^2\sigma_i^2\right)t^2/2\right)}, \end{aligned}$$

the mgf of the indicated normal distribution. □

Theorem 4.6.11 (Generalization of Lemma 4.2.7) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random vectors. Then $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent random vectors if and only if there exist functions $g_i(\mathbf{x}_i), i = 1, \ldots, n$, such that the joint pdf or pmf of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ can be written as*

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = g_1(\mathbf{x}_1)\cdot\cdots\cdot g_n(\mathbf{x}_n).$$

Theorem 4.6.12 (Generalization of Theorem 4.3.5) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors. Let $g_i(\mathbf{x}_i)$ be a function only of $\mathbf{x}_i, i = 1, \ldots, n$. Then the random variables $U_i = g_i(\mathbf{X}_i), i = 1, \ldots, n$, are mutually independent.*

We close this section by describing the generalization of a technique for finding the distribution of a transformation of a random vector. We will present the generalization of formula (4.3.6) that gives the pdf of the new random vector in terms of the pdf of the original random vector. Note that to fully understand the remainder of this section, some knowledge of matrix algebra is required. (See, for example, Searle 1982.) In particular, we will need to compute the determinant of a matrix. This is the only place in the book where such knowledge is required.

Let $(X_1, \ldots, X_n)$ be a random vector with pdf $f_{\mathbf{X}}(x_1, \ldots, x_n)$. Let $\mathcal{A} = \{\mathbf{x} : f_{\mathbf{X}}(\mathbf{x}) > 0\}$. Consider a new random vector $(U_1, \ldots, U_n)$, defined by $U_1 = g_1(X_1, \ldots, X_n), U_2 = g_2(X_1, \ldots, X_n), \ldots, U_n = g_n(X_1, \ldots, X_n)$. Suppose that $A_0, A_1, \ldots, A_k$ form a partition of $\mathcal{A}$ with these properties. The set A_0, which may be empty, satisfies $P((X_1, \ldots, X_n) \in A_0) = 0$. The transformation $(U_1, \ldots, U_n) = (g_1(\mathbf{X}), \ldots, g_n(\mathbf{X}))$ is a one-to-one transformation from A_i onto $\mathcal{B}$ for each $i = 1, 2, \ldots, k$. Then for each i, the inverse functions from $\mathcal{B}$ to A_i can be found. Denote the ith inverse by $x_1 = h_{1i}(u_1, \ldots, u_n)$, $x_2 = h_{2i}(u_1, \ldots, u_n), \ldots, x_n = h_{ni}(u_1, \ldots, u_n)$. This ith inverse gives, for $(u_1, \ldots, u_n) \in \mathcal{B}$, the unique $(x_1, \ldots, x_n) \in A_i$ such that $(u_1, \ldots, u_n) = (g_1(x_1, \ldots, x_n), \ldots, g_n(x_1, \ldots, x_n))$. Let J_i denote the Jacobian computed from the ith inverse. That is,

$$J_i = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{1i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_{1i}(\mathbf{u})}{\partial u_n} \\ \frac{\partial h_{2i}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_{2i}(\mathbf{u})}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_{ni}(\mathbf{u})}{\partial u_1} & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial h_{ni}(\mathbf{u})}{\partial u_n} \end{vmatrix},$$

the determinant of an $n \times n$ matrix. Assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the joint pdf, $f_{\mathbf{U}}(u_1, \ldots, u_n)$, for $\mathbf{u} \in \mathcal{B}$:

$$f_{\mathbf{U}}(u_1, \ldots, u_n) = \sum_{i=1}^{k} f_{\mathbf{X}}(h_{1i}(u_1, \ldots, u_n), \ldots, h_{ni}(u_1, \ldots, u_n))|J_i|. \tag{4.6.7}$$

Example 4.6.13 (Multivariate change of variables) Let (X_1, X_2, X_3, X_4) have joint pdf

$$f_{\mathbf{X}}(x_1, x_2, x_3, x_4) = 24e^{-x_1 - x_2 - x_3 - x_4}, \quad 0 < x_1 < x_2 < x_3 < x_4 < \infty.$$

Consider the transformation

$$U_1 = X_1, \quad U_2 = X_2 - X_1, \quad U_3 = X_3 - X_2, \quad U_4 = X_4 - X_3.$$

This transformation maps the set $\mathcal{A}$ onto the set $\mathcal{B} = \{\mathbf{u} : 0 < u_i < \infty, i = 1, 2, 3, 4\}$. The transformation is one-to-one, so $k = 1$, and the inverse is

$$X_1 = U_1, \quad X_2 = U_1 + U_2, \quad X_3 = U_1 + U_2 + U_3, \quad X_4 = U_1 + U_2 + U_3 + U_4.$$

The Jacobian of the inverse is

$$J = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{vmatrix} = 1.$$

Since the matrix is triangular, the determinant is equal to the product of the diagonal elements. Thus, from (4.6.7) we obtain

$$\begin{aligned} f_U(u_1,\ldots,u_4) &= 24e^{-u_1-(u_1+u_2)-(u_1+u_2+u_3)-(u_1+u_2+u_3+u_4)} \\ &= 24e^{-4u_1-3u_2-2\mu_3-u_4} \end{aligned}$$

on $\mathcal{B}$. From this the marginal pdfs of U_1, U_2, U_3, and U_4 can be calculated. It turns out that $f_U(u_i) = (5-i)e^{-(5-i)u_i}, 0 < u_i$; that is, $U_i \sim$ exponential$(1/(5-i))$. From Theorem 4.6.11 we see that U_1, U_2, U_3, and U_4 are mutually independent random variables. ||

The model in Example 4.6.13 can arise in the following way. Suppose Y_1, Y_2, Y_3, and Y_4 are mutually independent random variables, each with an exponential(1) distribution. Define $X_1 = \min(Y_1, Y_2, Y_3, Y_4)$, $X_2 =$ second smallest value of (Y_1, Y_2, Y_3, Y_4), $X_3 =$ second largest value of (Y_1, Y_2, Y_3, Y_4), and $X_4 = \max(Y_1, Y_2, Y_3, Y_4)$. These variables will be called *order statistics* in Section 5.5. There we will see that the joint pdf of (X_1, X_2, X_3, X_4) is the pdf given in Example 4.6.13. Now the variables U_2, U_3, and U_4 defined in the example are called the *spacings* between the order statistics. The example showed that, for these exponential random variables $(Y_1, \ldots, Y_n)$, the spacings between the order statistics are mutually independent and also have exponential distributions.

4.7 Inequalities

In Section 3.6 we saw inequalities that were derived using probabilistic arguments. In this section we will see inequalities that apply to probabilities and expectations but are based on arguments that use properties of functions and numbers.

4.7.1 Numerical Inequalities

The inequalities in this subsection, although often stated in terms of expectations, rely mainly on properties of numbers. In fact, they are all based on the following simple lemma.

Lemma 4.7.1 *Let a and b be any positive numbers, and let p and q be any positive numbers (necessarily greater than 1) satisfying*

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{4.7.1}$$

Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab \tag{4.7.2}$$

with equality if and only if $a^p = b^q$.

Proof: Fix b, and consider the function

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab.$$

To minimize $g(a)$, differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \Rightarrow a^{p-1} - b = 0 \Rightarrow b = a^{p-1}.$$

A check of the second derivative will establish that this is indeed a minimum. The value of the function at the minimum is

$$\begin{aligned} \frac{1}{p}a^p + \frac{1}{q}\left(a^{p-1}\right)^q - aa^{p-1} &= \frac{1}{p}a^p + \frac{1}{q}a^p - a^p \quad \begin{pmatrix}(p-1)q = p \text{ follows} \\ \text{from (4.7.1)}\end{pmatrix} \\ &= 0. \qquad \text{(again from (4.7.1))} \end{aligned}$$

Hence the minimum is 0 and (4.7.2) is established. Since the minimum is unique (why?), equality holds only if $a^{p-1} = b$, which is equivalent to $a^p = b^q$, again from (4.7.1). □

The first of our expectation inequalities, one of the most used and most important, follows easily from the lemma.

Theorem 4.7.2 (Hölder's Inequality) *Let X and Y be any two random variables, and let p and q satisfy (4.7.1). Then*

$$|\mathrm{E}XY| \le \mathrm{E}|XY| \le (\mathrm{E}|X|^p)^{1/p}\,(\mathrm{E}|Y|^q)^{1/q}. \tag{4.7.3}$$

Proof: The first inequality follows from $-|XY| \le XY \le |XY|$ and Theorem 2.2.5. To prove the second inequality, define

$$a = \frac{|X|}{(\mathrm{E}|X|^p)^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{(\mathrm{E}|Y|^q)^{1/q}}.$$

Applying Lemma 4.7.1, we get

$$\frac{1}{p}\frac{|X|^p}{\mathrm{E}\,|X|^p} + \frac{1}{q}\frac{|Y|^q}{\mathrm{E}\,|Y|^q} \ge \frac{|XY|}{(\mathrm{E}\,|X|^p)^{1/p}\,(\mathrm{E}|Y|^q)^{1/q}}.$$

Now take expectations of both sides. The expectation of the left-hand side is 1, and rearrangement gives (4.7.3). □

Perhaps the most famous special case of Hölder's Inequality is that for which $p = q = 2$. This is called the Cauchy–Schwarz Inequality.

Theorem 4.7.3 (Cauchy–Schwarz Inequality) *For any two random variables X and Y,*

$$|\mathrm{E}XY| \le \mathrm{E}|XY| \le (\mathrm{E}|X|^2)^{1/2}(\mathrm{E}|Y|^2)^{1/2}. \tag{4.7.4}$$

Example 4.7.4 (Covariance inequality) If X and Y have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively, we can apply the Cauchy–Schwarz Inequality to get

$$\mathrm{E}\,|(X-\mu_X)(Y-\mu_Y)| \le \left\{\mathrm{E}(X-\mu_X)^2\right\}^{1/2}\left\{\mathrm{E}(Y-\mu_Y)^2\right\}^{1/2}.$$

Squaring both sides and using statistical notation, we have

$$(\mathrm{Cov}(X,Y))^2 \le \sigma_X^2\sigma_Y^2.$$

Recalling the definition of the correlation coefficient, ρ, we have proved that $0 \le \rho^2 \le 1$. Furthermore, the condition for equality in Lemma 4.7.1 still carries over, and equality is attained here only if $X-\mu_X = c(Y-\mu_Y)$, for some constant c. That is, the correlation is ± 1 if and only if X and Y are linearly related. Compare the ease of this proof to the one used in Theorem 4.5.7, before we had the Cauchy–Schwarz Inequality. ‖

Some other special cases of Hölder's Inequality are often useful. If we set $Y \equiv 1$ in (4.7.3), we get

$$\mathrm{E}|X| \le \{\mathrm{E}(|X|^p)\}^{1/p}, \quad 1 < p < \infty. \tag{4.7.5}$$

For $1 < r < p$, if we replace $|X|$ by $|X|^r$ in (4.7.5), we obtain

$$\mathrm{E}|X|^r \le \{\mathrm{E}(|X|^{pr})\}^{1/p}.$$

Now write $s = pr$ (note that $s > r$) and rearrange terms to get

$$\{\mathrm{E}|X|^r\}^{1/r} \le \{\mathrm{E}|X|^s\}^{1/s}, \quad 1 < r < s < \infty, \tag{4.7.6}$$

which is known as *Liapounov's Inequality.*

Our next named inequality is similar in spirit to Hölder's Inequality and, in fact, follows from it.

Theorem 4.7.5 (Minkowski's Inequality) *Let X and Y be any two random variables. Then for $1 \le p < \infty$,*

$$[\mathrm{E}|X+Y|^p]^{1/p} \le [\mathrm{E}|X|^p]^{1/p} + [\mathrm{E}|Y|^p]^{1/p}. \tag{4.7.7}$$

Proof: Write

$$\begin{aligned}\mathrm{E}|X+Y|^p &= \mathrm{E}\left(|X+Y||X+Y|^{p-1}\right)\\ &\le \mathrm{E}\left(|X||X+Y|^{p-1}\right) + \mathrm{E}\left(|Y||X+Y|^{p-1}\right),\end{aligned} \tag{4.7.8}$$

where we have used the fact that $|X+Y| \le |X| + |Y|$ (the *triangle inequality*; see Exercise 4.64). Now apply Hölder's Inequality to each expectation on the right-hand side of (4.7.8) to get

$$\begin{aligned}\mathrm{E}(|X+Y|^p) \le{}& \left[\mathrm{E}(|X|^p)\right]^{1/p}\left[\mathrm{E}|X+Y|^{q(p-1)}\right]^{1/q}\\ &+ \left[\mathrm{E}(|Y|^p)\right]^{1/p}\left[\mathrm{E}|X+Y|^{q(p-1)}\right]^{1/q},\end{aligned}$$

where q satisfies $1/p+1/q=1$. Now divide through by $\left[\mathrm{E}(|X+Y|^{q(p-1)})\right]^{1/q}$. Noting that $q(p-1)=p$ and $1-1/q=1/p$, we obtain (4.7.7). □

The preceding theorems also apply to numerical sums where there is no explicit reference to an expectation. For example, for numbers $a_i, b_i, i=1,\ldots,n$, the inequality

$$\sum_{i=1}^{n}|a_ib_i| \le \left(\sum_{i=1}^{n}a_i^p\right)^{1/p}\left(\sum_{i=1}^{n}b_i^q\right)^{1/q}, \quad \frac{1}{p}+\frac{1}{q}=1, \tag{4.7.9}$$

is a version of Hölder's Inequality. To establish (4.7.9) we can formally set up an expectation with respect to random variables taking values $a_1,\ldots,a_n$ and $b_1,\ldots,b_n$. (This is done in Example 4.7.8.)

An important special case of (4.7.9) occurs when $b_i \equiv 1, p=q=2$. We then have

$$\frac{1}{n}\left(\sum_{i=1}^{n}|a_i|\right)^2 \le \sum_{i=1}^{n}a_i^2.$$

4.7.2 Functional Inequalities

The inequalities in this section rely on properties of real-valued functions rather than on any statistical properties. In many cases, however, they prove to be very useful. One of the most useful is Jensen's Inequality, which applies to convex functions.

Definition 4.7.6 A function $g(x)$ is *convex* if $g(\lambda x+(1-\lambda)y) \le \lambda g(x)+(1-\lambda)g(y)$, for all x and y, and $0<\lambda<1$. The function $g(x)$ is *concave* if $-g(x)$ is convex.

Informally, we can think of convex functions as functions that "hold water"—that is, they are bowl-shaped ($g(x)=x^2$ is convex), while concave functions "spill water" ($g(x)=\log x$ is concave). More formally, convex functions lie below lines connecting any two points (see Figure 4.7.1). As λ goes from 0 to 1, $\lambda g(x_1)+(1-\lambda)g(x_2)$

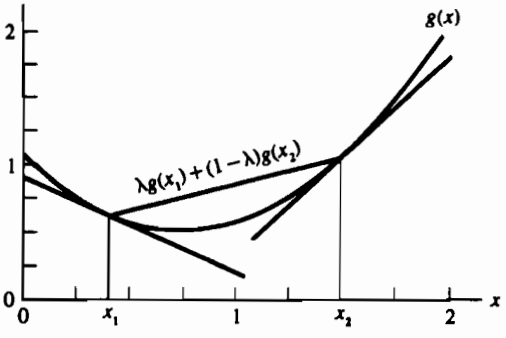

Figure 4.7.1. *Convex function and tangent lines at x_1 and x_2*

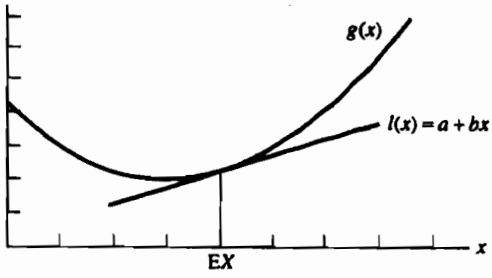

Figure 4.7.2. *Graphical illustration of Jensen's Inequality*

defines a line connecting $g(x_1)$ and $g(x_2)$. This line lies above $g(x)$ if $g(x)$ is convex. Furthermore, a convex function lies above all of its tangent lines (also shown in Figure 4.7.1), and that fact is the basis of Jensen's Inequality.

Theorem 4.7.7 (Jensen's Inequality) *For any random variable X, if $g(x)$ is a convex function, then*

$$\mathrm{E}g(X) \geq g(\mathrm{E}X).$$

Equality holds if and only if, for every line $a + bx$ that is tangent to $g(x)$ at $x = EX$, $P(g(X) = a + bX) = 1$.

Proof: To establish the inequality, let $l(x)$ be a tangent line to $g(x)$ at the point $g(\mathrm{E}X)$. (Recall that $\mathrm{E}X$ is a constant.) Write $l(x) = a + bx$ for some a and b. The situation is illustrated in Figure 4.7.2.

Now, by the convexity of g we have $g(x) \geq a + bx$. Since expectations preserve inequalities,

$$\begin{aligned}
\mathrm{E}g(X) &\geq \mathrm{E}(a + bX) \\
&= a + b\mathrm{E}X && \left(\begin{array}{c}\text{linearity of expectation,}\\ \text{Theorem 2.2.5}\end{array}\right) \\
&= l(\mathrm{E}X) && (\text{definition of } l(x)) \\
&= g(\mathrm{E}X), && (l \text{ is tangent at } \mathrm{E}X)
\end{aligned}$$

as was to be shown.

If $g(x)$ is linear, equality follows from properties of expectations (Theorem 2.2.5). For the "only if" part see Exercise 4.62. □

One immediate application of Jensen's Inequality shows that $\mathrm{E}X^2 \geq (\mathrm{E}X)^2$, since $g(x) = x^2$ is convex. Also, if x is positive, then $1/x$ is convex; hence $\mathrm{E}(1/X) \geq 1/\mathrm{E}X$, another useful application.

To check convexity of a twice differentiable function is quite easy. The function $g(x)$ is convex if $g''(x) \geq 0$, for all x, and $g(x)$ is concave if $g''(x) \leq 0$, for all x. Jensen's Inequality applies to concave functions as well. If g is concave, then $\mathrm{E}g(X) \leq g(\mathrm{E}X)$.

Example 4.7.8 (An inequality for means) Jensen's Inequality can be used to prove an inequality between three different kinds of means. If $a_1, \ldots, a_n$ are positive numbers, define

$$a_A = \frac{1}{n}(a_1 + a_2 + \cdots + a_n), \qquad \text{(arithmetic mean)}$$

$$a_G = [a_1 a_2 \cdot \cdots \cdot a_n]^{1/n}, \qquad \text{(geometric mean)}$$

$$a_H = \frac{1}{\frac{1}{n}\left(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n}\right)}. \qquad \text{(harmonic mean)}$$

An inequality relating these means is

$$a_H \le a_G \le a_A.$$

To apply Jensen's Inequality, let X be a random variable with range $a_1, \ldots, a_n$ and $P(X = a_i) = 1/n, i = 1, \ldots, n$. Since $\log x$ is a concave function, Jensen's Inequality shows that $\mathrm{E}(\log X) \le \log(\mathrm{E}X)$; hence,

$$\log a_G = \frac{1}{n}\sum_{i=1}^{n} \log a_i = \mathrm{E}(\log X) \le \log(\mathrm{E}X) = \log\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) = \log a_A,$$

so $a_G \le a_A$. Now again use the fact that $\log x$ is concave to get

$$\log \frac{1}{a_H} = \log\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_i}\right) = \log \mathrm{E}\frac{1}{X} \ge \mathrm{E}\left(\log \frac{1}{X}\right) = -\mathrm{E}(\log X).$$

Since $\mathrm{E}(\log X) = \log a_G$, it then follows that $\log(1/a_H) \ge \log(1/a_G)$, or $a_G \ge a_H$. ||

The next inequality merely exploits the definition of covariance, but sometimes proves to be useful. If X is a random variable with finite mean μ and $g(x)$ is a nondecreasing function, then

$$\mathrm{E}\left(g(X)(X - \mu)\right) \ge 0,$$

since

$$\begin{aligned}
&\mathrm{E}(g(X)(X-\mu)) \\
&\quad = \mathrm{E}\left(g(X)(X-\mu)I_{(-\infty,0)}(X-\mu)\right) + \mathrm{E}\left(g(X)(X-\mu)I_{[0,\infty)}(X-\mu)\right) \\
&\quad \ge \mathrm{E}\left(g(\mu)(X-\mu)I_{(-\infty,0)}(X-\mu)\right) \\
&\qquad + \mathrm{E}\left(g(\mu)(X-\mu)I_{[0,\infty)}(X-\mu)\right) \qquad \text{(since } g \text{ is nondecreasing)} \\
&\quad = g(\mu)\mathrm{E}(X-\mu) \\
&\quad = 0.
\end{aligned}$$

A generalization of this argument can be used to establish the following inequality (see Exercise 4.65).

Theorem 4.7.9 (Covariance Inequality) *Let X be any random variable and $g(x)$ and $h(x)$ any functions such that $\mathrm{E}g(X)$, $\mathrm{E}h(X)$, and $\mathrm{E}(g(X)h(X))$ exist.*

a. *If $g(x)$ is a nondecreasing function and $h(x)$ is a nonincreasing function, then*

$$\mathrm{E}\left(g(X)h(X)\right) \leq \left(\mathrm{E}g(X)\right)\left(\mathrm{E}h(X)\right).$$

b. *If $g(x)$ and $h(x)$ are either both nondecreasing or both nonincreasing, then*

$$\mathrm{E}\left(g(X)h(X)\right) \geq \left(\mathrm{E}g(X)\right)\left(\mathrm{E}h(X)\right).$$

The intuition behind the inequality is easy. In case (a) there is negative correlation between g and h, while in case (b) there is positive correlation. The inequalities merely reflect this fact. The usefulness of the Covariance Inequality is that it allows us to bound an expectation without using higher-order moments.

4.8 Exercises

4.1 A random point (X, Y) is distributed uniformly on the square with vertices $(1, 1)$, $(1, -1)$, $(-1, 1)$, and $(-1, -1)$. That is, the joint pdf is $f(x, y) = \frac{1}{4}$ on the square. Determine the probabilities of the following events.

(a) $X^2 + Y^2 < 1$
(b) $2X - Y > 0$
(c) $|X + Y| < 2$

4.2 Prove the following properties of bivariate expectations (the bivariate analog to Theorem 2.2.5). For random variables X and Y, functions $g_1(x, y)$ and $g_2(x, y)$, and constants a, b, and c:

(a) $\mathrm{E}(ag_1(X, Y) + bg_2(X, Y) + c) = a\mathrm{E}(g_1(X, Y)) + b\mathrm{E}(g_2(X, Y)) + c$.
(b) If $g_1(x, y) \geq 0$, then $\mathrm{E}(g_1(X, Y)) \geq 0$.
(c) If $g_1(x, y) \geq g_2(x, y)$, then $\mathrm{E}(g_1(X, Y)) \geq \mathrm{E}(g_2(X, Y))$.
(d) If $a \leq g_1(x, y) \leq b$, then $a \leq \mathrm{E}(g_1(X, Y)) \leq b$.

4.3 Using Definition 4.1.1, show that the random vector (X, Y) defined at the end of Example 4.1.5 has the pmf given in that example.

4.4 A pdf is defined by

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

(a) Find the value of C.
(b) Find the marginal distribution of X.
(c) Find the joint cdf of X and Y.
(d) Find the pdf of the random variable $Z = 9/(X + 1)^2$.

4.5 (a) Find $P(X > \sqrt{Y})$ if X and Y are jointly distributed with pdf

$$f(x, y) = x + y, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

(b) Find $P(X^2 < Y < X)$ if X and Y are jointly distributed with pdf

$$f(x, y) = 2x, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1.$$

4.6 A and B agree to meet at a certain place between 1 PM and 2 PM. Suppose they arrive at the meeting place independently and randomly during the hour. Find the distribution of the length of time that A waits for B. (If B arrives before A, define A's waiting time as 0.)

4.7 A woman leaves for work between 8 AM and 8:30 AM and takes between 40 and 50 minutes to get there. Let the random variable X denote her time of departure, and the random variable Y the travel time. Assuming that these variables are independent and uniformly distributed, find the probability that the woman arrives at work before 9 AM.

4.8 Referring to Miscellanea 4.9.1.

(a) Show that $P(X = m|M = m) = P(X = 2m|M = m) = 1/2$, and verify the expressions for $P(M = x|X = x)$ and $P(M = x/2|X = x)$.

(b) Verify that one should trade only if $\pi(x/2) < 2\pi(x)$, and if π is the exponential(λ) density, show that it is optimal to trade if $x < 2\log 2/\lambda$.

(c) For the classical approach, show that $P(Y = 2x|X = m) = 1$ and $P(Y = x/2|X = 2m) = 1$ and that your expected winning if you trade or keep your envelope is $E(Y) = 3m/2$.

4.9 Prove that if the joint cdf of X and Y satisfies

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

then for any pair of intervals (a, b), and (c, d),

$$P(a \le X \le b, c \le Y \le d) = P(a \le X \le b)P(c \le Y \le d).$$

4.10 The random pair (X, Y) has the distribution

		X		
		1	2	3
Y	2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
	3	$\frac{1}{6}$	0	$\frac{1}{6}$
	4	0	$\frac{1}{3}$	0

(a) Show that X and Y are dependent.

(b) Give a probability table for random variables U and V that have the same marginals as X and Y but are independent.

4.11 Let U = the number of trials needed to get the first head and V = the number of trials needed to get two heads in repeated tosses of a fair coin. Are U and V independent random variables?

4.12 If a stick is broken at random into three pieces, what is the probability that the pieces can be put together in a triangle? (See Gardner 1961 for a complete discussion of this problem.)

4.13 Let X and Y be random variables with finite means.

(a) Show that

$$\min_{g(x)} \mathrm{E}\,(Y - g(X))^2 = \mathrm{E}\,(Y - \mathrm{E}(Y|X))^2,$$

where $g(x)$ ranges over all functions. $\mathrm{E}(Y|X)$ is sometimes called the *regression of Y on X*, the "best" predictor of Y conditional on X.)

(b) Show that equation (2.2.4) can be derived as a special case of part (a).

4.14 Suppose X and Y are independent n$(0,1)$ random variables.

(a) Find $P(X^2+Y^2<1)$.

(b) Find $P(X^2<1)$, after verifying that X^2 is distributed χ^2_1.

4.15 Let $X \sim$ Poisson(θ), $Y \sim$ Poisson(λ), independent. It was shown in Theorem 4.3.2 that the distribution of $X+Y$ is Poisson$(\theta+\lambda)$. Show that the distribution of $X|X+Y$ is binomial with success probability $\theta/(\theta+\lambda)$. What is the distribution of $Y|X+Y$?

4.16 Let X and Y be independent random variables with the same geometric distribution.

(a) Show that U and V are independent, where U and V are defined by

$$U = \min(X,Y) \quad \text{and} \quad V = X-Y.$$

(b) Find the distribution of $Z = X/(X+Y)$, where we define $Z=0$ if $X+Y=0$.

(c) Find the joint pdf of X and $X+Y$.

4.17 Let X be an exponential(1) random variable, and define Y to be the integer part of $X+1$, that is

$$Y = i+1 \quad \text{if and only if} \quad i \le X < i+1, \quad i = 0,1,2,\ldots.$$

(a) Find the distribution of Y. What well-known distribution does Y have?

(b) Find the conditional distribution of $X-4$ given $Y \ge 5$.

4.18 Given that $g(x) \ge 0$ has the property that

$$\int_0^\infty g(x)\,dx = 1,$$

show that

$$f(x,y) = \frac{2g\left(\sqrt{x^2+y^2}\right)}{\pi\sqrt{x^2+y^2}}, \quad x,y>0,$$

is a pdf.

4.19 (a) Let X_1 and X_2 be independent n$(0,1)$ random variables. Find the pdf of $(X_1-X_2)^2/2$.

(b) If $X_i, i=1,2$, are independent gamma$(\alpha_i,1)$ random variables, find the marginal distributions of $X_1/(X_1+X_2)$ and $X_2/(X_1+X_2)$.

4.20 X_1 and X_2 are independent n$(0,\sigma^2)$ random variables.

(a) Find the joint distribution of Y_1 and Y_2, where

$$Y_1 = X_1^2+X_2^2 \quad \text{and} \quad Y_2 = \frac{X_1}{\sqrt{Y_1}}.$$

(b) Show that Y_1 and Y_2 are independent, and interpret this result geometrically.

4.21 A point is generated at random in the plane according to the following polar scheme. A radius R is chosen, where the distribution of R^2 is χ^2 with 2 degrees of freedom. Independently, an angle θ is chosen, where $\theta \sim$ uniform$(0,2\pi)$. Find the joint distribution of $X = R\cos\theta$ and $Y = R\sin\theta$.

4.22 Let (X, Y) be a bivariate random vector with joint pdf $f(x, y)$. Let $U = aX + b$ and $V = cY + d$, where a, b, c, and d are fixed constants with $a > 0$ and $c > 0$. Show that the joint pdf of (U, V) is

$$f_{U,V}(u, v) = \frac{1}{ac} f\left(\frac{u-b}{a}, \frac{v-d}{c}\right).$$

4.23 For X and Y as in Example 4.3.3, find the distribution of XY by making the transformations given in (a) and (b) and integrating out V.

(a) $U = XY$, $V = Y$

(b) $U = XY$, $V = X/Y$

4.24 Let X and Y be independent random variables with $X \sim$ gamma$(r, 1)$ and $Y \sim$ gamma$(s, 1)$. Show that $Z_1 = X + Y$ and $Z_2 = X/(X + Y)$ are independent, and find the distribution of each. (Z_1 is gamma and Z_2 is beta.)

4.25 Use the techniques of Section 4.3 to derive the joint distribution of (X, Y) from the joint distribution of (X, Z) in Examples 4.5.8 and 4.5.9.

4.26 X and Y are independent random variables with $X \sim$ exponential(λ) and $Y \sim$ exponential(μ). It is impossible to obtain direct observations of X and Y. Instead, we observe the random variables Z and W, where

$$Z = \min\{X, Y\} \qquad \text{and} \qquad W = \begin{cases} 1 & \text{if } Z = X \\ 0 & \text{if } Z = Y. \end{cases}$$

(This is a situation that arises, in particular, in medical experiments. The X and Y variables are *censored.*)

(a) Find the joint distribution of Z and W.

(b) Prove that Z and W are independent. (*Hint:* Show that $P(Z \le z | W = i) = P(Z \le z)$ for $i = 0$ or 1.)

4.27 Let $X \sim \text{n}(\mu, \sigma^2)$ and let $Y \sim \text{n}(\gamma, \sigma^2)$. Suppose X and Y are independent. Define $U = X + Y$ and $V = X - Y$. Show that U and V are independent normal random variables. Find the distribution of each of them.

4.28 Let X and Y be independent standard normal random variables.

(a) Show that $X/(X + Y)$ has a Cauchy distribution.

(b) Find the distribution of $X/|Y|$.

(c) Is the answer to part (b) surprising? Can you formulate a general theorem?

4.29 Jones (1999) looked at the distribution of functions of X and Y when $X = R\cos\theta$ and $Y = R\sin\theta$, where $\theta \sim U(0, 2\pi)$ and R is a positive random variable. Here are two of the many situations that he considered.

(a) Show that X/Y has a Cauchy distribution.

(b) Show that the distribution of $(2XY)/\sqrt{X^2 + Y^2}$ is the same as the distribution of X. Specialize this result to one about $\text{n}(0, \sigma^2)$ random variables.

4.30 Suppose the distribution of Y, conditional on $X = x$, is $\text{n}(x, x^2)$ and that the marginal distribution of X is uniform$(0, 1)$.

(a) Find $\text{E}Y$, $\text{Var}\,Y$, and $\text{Cov}(X, Y)$.

(b) Prove that Y/X and X are independent.

4.31 Suppose that the random variable Y has a binomial distribution with n trials and success probability X, where n is a given constant and X is a uniform$(0,1)$ random variable.

(a) Find EY and $\text{Var}\, Y$.
(b) Find the joint distribution of X and Y.
(c) Find the marginal distribution of Y.

4.32 (a) For the hierarchical model

$$Y|\Lambda \sim \text{Poisson}(\Lambda) \quad \text{and} \quad \Lambda \sim \text{gamma}(\alpha, \beta)$$

find the marginal distribution, mean, and variance of Y. Show that the marginal distribution of Y is a negative binomial if α is an integer.

(b) Show that the three-stage model

$$Y|N \sim \text{binomial}(N, p), \quad N|\Lambda \sim \text{Poisson}(\Lambda), \quad \text{and} \quad \Lambda \sim \text{gamma}(\alpha, \beta)$$

leads to the same marginal (unconditional) distribution of Y.

4.33 (*Alternative derivation of the negative binomial distribution*) Solomon (1983) details the following biological model. Suppose that each of a random number, N, of insects lays X_i eggs, where the X_is are independent, identically distributed random variables. The total number of eggs laid is $H = X_1 + \cdots + X_N$. What is the distribution of H? It is common to assume that N is Poisson(λ). Furthermore, if we assume that each X_i has the logarithmic series distribution (see Exercise 3.14) with success probability p, we have the hierarchical model

$$H|N = X_1 + \cdots + X_N, \quad P(X_i = t) = \frac{-1}{\log(p)} \frac{(1-p)^t}{t},$$

$$N \sim \text{Poisson}(\lambda).$$

Show that the marginal distribution of H is negative binomial(r, p), where $r = -\lambda/\log(p)$. (It is easiest to calculate and identify the mgf of H using Theorems 4.4.3 and 4.6.7. Stuart and Ord 1987, Section 5.21, also mention this derivation of the logarithmic series distribution. They refer to H as a *randomly stopped sum*.)

4.34 (a) For the hierarchy in Example 4.4.6, show that the marginal distribution of X is given by the *beta-binomial distribution*,

$$P(X = x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}.$$

(b) A variation on the hierarchical model in part (a) is

$$X|P \sim \text{negative binomial}(r, P) \quad \text{and} \quad P \sim \text{beta}(\alpha, \beta).$$

Find the marginal pmf of X and its mean and variance. (This distribution is the *beta-Pascal*.)

4.35 (a) For the hierarchy in Example 4.4.6, show that the variance of X can be written

$$\text{Var}\, X = nEP(1 - EP) + n(n-1)\,\text{Var}\, P.$$

(The first term reflects binomial variation with success probability EP, and the second term is often called "extra-binomial" variation, showing how the hierarchical model has a variance that is larger than the binomial alone.)

(b) For the hierarchy in Exercise 4.32, show that the variance of Y can be written

$$\operatorname{Var} Y = \mathrm{E}\Lambda + \operatorname{Var}\Lambda = \mu + \frac{1}{\alpha}\mu^2,$$

where $\mu = \mathrm{E}\Lambda$. Identify the "extra-Poisson" variation induced by the hierarchy.

4.36 One generalization of the Bernoulli trials hierarchy in Example 4.4.6 is to allow the success probability to vary from trial to trial, keeping the trials independent. A standard model for this situation is

$$X_i|P_i \sim \text{Bernoulli}(P_i), \quad i = 1, \ldots, n,$$
$$P_i \sim \text{beta}(\alpha, \beta).$$

This model might be appropriate, for example, if we are measuring the success of a drug on n patients and, because the patients are different, we are reluctant to assume that the success probabilities are constant. (This can be thought of as an *empirical Bayes model*; see Miscellanea 7.5.6.)

A random variable of interest is $Y = \sum_{i=1}^{n} X_i$, the total number of successes.

(a) Show that $\mathrm{E}Y = n\alpha/(\alpha+\beta)$.

(b) Show that $\operatorname{Var} Y = n\alpha\beta/(\alpha+\beta)^2$, and hence Y has the same mean and variance as a binomial$(n, \frac{\alpha}{\alpha+\beta})$ random variable. What is the distribution of Y?

(c) Suppose now that the model is

$$X_i|P_i \sim \text{binomial}(n_i, P_i), \quad i = 1, \ldots, k,$$
$$P_i \sim \text{beta}(\alpha, \beta).$$

Show that for $Y = \sum_{i=1}^{k} X_i$, $\mathrm{E}Y = \frac{\alpha}{\alpha+\beta}\sum_{i=1}^{k} n_i$ and $\operatorname{Var} Y = \sum_{i=1}^{k} \operatorname{Var} X_i$, where

$$\operatorname{Var} X_i = n_i \frac{\alpha\beta(\alpha+\beta+n_i)}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

4.37 A generalization of the hierarchy in Exercise 4.34 is described by D. G. Morrison (1978), who gives a model for *forced binary choices*. A forced binary choice occurs when a person is forced to choose between two alternatives, as in a taste test. It may be that a person cannot actually discriminate between the two choices (can you tell Coke from Pepsi?), but the setup of the experiment is such that a choice must be made. Therefore, there is a confounding between discriminating correctly and guessing correctly. Morrison modeled this by defining the following parameters:

p = probability that a person can actually discriminate,

c = probability that a person discriminates correctly.

Then

$$c = p + \frac{1}{2}(1-p) = \frac{1}{2}(1+p), \quad \frac{1}{2} < c < 1,$$

where $\frac{1}{2}(1-p)$ is the probability that a person guesses correctly. We now run the experiment and observe $X_1, \ldots, X_n \sim \text{Bernoulli}(c)$, so

$$P(\Sigma X_i = k|c) = \binom{n}{k} c^k (1-c)^{n-k}.$$

However, it is probably the case that p is not constant from person to person, so p is allowed to vary according to a beta distribution,

$$P \sim \text{beta}(a, b).$$

(a) Show that the distribution of ΣX_i is beta-binomial.
(b) Find the mean and variance of ΣX_i.

4.38 (*The gamma as a mixture of exponentials*) Gleser (1989) shows that, in certain cases, the gamma distribution can be written as a scale mixture of exponentials, an identity suggested by different analyses of the same data. Let $f(x)$ be a gamma(r, λ) pdf.

(a) Show that if $r \leq 1$, then $f(x)$ can be written

$$f(x) = \int_0^\lambda \frac{1}{\nu} e^{-x/\nu} p_\lambda(\nu)\, d\nu,$$

where

$$p_\lambda(\nu) = \frac{1}{\Gamma(r)\Gamma(1-r)} \frac{\nu^{r-1}}{(\lambda-\nu)^r}, \quad 0 < \nu < \lambda.$$

(*Hint:* Make a change of variable from ν to u, where $u = x/\nu - x/\lambda$.)
(b) Show that $p_\lambda(\nu)$ is a pdf, for $r \leq 1$, by showing that

$$\int_0^\lambda p_\lambda(\nu)\, d\nu = 1.$$

(c) Show that the restriction $r \leq 1$ is necessary for the representation in part (a) to be valid; that is, there is no such representation if $r > 1$. (*Hint:* Suppose $f(x)$ can be written $f(x) = \int_0^\infty (e^{-x/\nu}/\nu) q_\lambda(\nu) d\nu$ for some pdf $q_\lambda(\nu)$. Show that $\frac{\partial}{\partial x} \log(f(x)) > 0$ but $\frac{\partial}{\partial x} \log\left(\int_0^\infty (e^{-x/\nu}/\nu) q_\lambda(\nu) d\nu\right) < 0$, a contradiction.)

4.39 Let $(X_1, \ldots, X_n)$ have a multinomial distribution with m trials and cell probabilities $p_1, \ldots, p_n$ (see Definition 4.6.2). Show that, for every i and j,

$$X_i | X_j = x_j \sim \text{binomial}\left(m - x_j, \frac{p_i}{1-p_j}\right)$$

$$X_j \sim \text{binomial}\,(m, p_j)$$

and that $\text{Cov}(X_i, X_j) = -mp_ip_j$.

4.40 A generalization of the beta distribution is the *Dirichlet* distribution. In its bivariate version, (X, Y) have pdf

$$f(x, y) = Cx^{a-1}y^{b-1}(1-x-y)^{c-1}, \quad 0 < x < 1, \quad 0 < y < 1, \quad 0 < y < 1 - x < 1,$$

where $a > 0$, $b > 0$, and $c > 0$ are constants.

(a) Show that $C = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)}$.
(b) Show that, marginally, both X and Y are beta.
(c) Find the conditional distribution of $Y|X = x$, and show that $Y/(1-x)$ is beta(b, c).
(d) Show that $\text{E}(XY) = \frac{ab}{(a+b+c+1)(a+b+c)}$, and find their covariance.

4.41 Show that any random variable is uncorrelated with a constant.

4.42 Let X and Y be independent random variables with means μ_X, μ_Y and variances σ_X^2, σ_Y^2. Find an expression for the correlation of XY and Y in terms of these means and variances.

4.43 Let X_1, X_2, and X_3 be uncorrelated random variables, each with mean μ and variance σ^2. Find, in terms of μ and σ^2, $\text{Cov}(X_1 + X_2, X_2 + X_3)$ and $\text{Cov}(X_1 + X_2, X_1 - X_2)$.

4.44 Prove the following generalization of Theorem 4.5.6: For any random vector $(X_1, \ldots, X_n)$,

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var } X_i + 2 \sum_{1 \le i < j \le n} \text{Cov}(X_i, X_j).$$

4.45 Show that if $(X, Y) \sim$ bivariate normal$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then the following are true.

(a) The marginal distribution of X is $\text{n}(\mu_X, \sigma_X^2)$ and the marginal distribution of Y is $\text{n}(\mu_Y, \sigma_Y^2)$.

(b) The conditional distribution of Y given $X = x$ is

$$\text{n}(\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X), \sigma_Y^2(1 - \rho^2)).$$

(c) For any constants a and b, the distribution of $aX + bY$ is

$$\text{n}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y).$$

4.46 (*A derivation of the bivariate normal distribution*) Let Z_1 and Z_2 be independent $\text{n}(0, 1)$ random variables, and define new random variables X and Y by

$$X = a_X Z_1 + b_X Z_2 + c_X \quad \text{and} \quad Y = a_Y Z_1 + b_Y Z_2 + c_Y,$$

where a_X, b_X, c_X, a_Y, b_Y, and c_Y are constants.

(a) Show that

$$\begin{aligned} \text{E}X &= c_X, \quad \text{Var } X = a_X^2 + b_X^2, \\ \text{E}Y &= c_Y, \quad \text{Var } Y = a_Y^2 + b_Y^2, \\ \text{Cov}(X, Y) &= a_X a_Y + b_X b_Y. \end{aligned}$$

(b) If we define the constants a_X, b_X, c_X, a_Y, b_Y, and c_Y by

$$a_X = \sqrt{\frac{1+\rho}{2}}\sigma_X, \quad b_X = \sqrt{\frac{1-\rho}{2}}\sigma_X, \quad c_X = \mu_X,$$

$$a_Y = \sqrt{\frac{1+\rho}{2}}\sigma_Y, \quad b_Y = -\sqrt{\frac{1-\rho}{2}}\sigma_Y, \quad c_Y = \mu_Y,$$

where $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ are constants, $-1 \le \rho \le 1$, then show that

$$\begin{aligned} \text{E}X &= \mu_X, \quad \text{Var } X = \sigma_X^2, \\ \text{E}Y &= \mu_Y, \quad \text{Var } Y = \sigma_Y^2, \\ \rho_{XY} &= \rho. \end{aligned}$$

(c) Show that (X, Y) has the bivariate normal pdf with parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ.

(d) If we start with bivariate normal parameters $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and ρ, we can define constants a_X, b_X, c_X, a_Y, b_Y, and c_Y as the solutions to the equations

$$\begin{aligned} \mu_X &= c_X, \quad \sigma_X^2 = a_X^2 + b_X^2, \\ \mu_Y &= c_Y, \quad \sigma_Y^2 = a_Y^2 + b_Y^2, \\ \rho\sigma_X\sigma_Y &= a_X a_Y + b_X b_Y. \end{aligned}$$

Show that the solution given in part (b) is not unique by exhibiting another solution to these equations. How many solutions are there?

4.47 (*Marginal normality does not imply bivariate normality.*) Let X and Y be independent n(0, 1) random variables, and define a new random variable Z by

$$Z = \begin{cases} X & \text{if } XY > 0 \\ -X & \text{if } XY < 0. \end{cases}$$

(a) Show that Z has a normal distribution.

(b) Show that the joint distribution of Z and Y is not bivariate normal. (*Hint*: Show that Z and Y always have the same sign.)

4.48 Gelman and Meng (1991) give an example of a bivariate family of distributions that are not bivariate normal but have normal conditionals. Define the joint pdf of (X, Y) as

$$f(x, y) \propto \exp\left\{-\frac{1}{2}\left[Ax^2y^2 + x^2 + y^2 - 2Bxy - 2Cx - 2Dy\right]\right\},$$

where A, B, C, D are constants.

(a) Show that the distribution of $X|Y = y$ is normal with mean $\frac{By+C}{Ay^2+1}$ and variance $\frac{1}{Ay^2+1}$. Derive a corresponding result for the distribution of $Y|X = x$.

(b) A most interesting configuration is $A = 1, B = 0, C = D = 8$. Show that this joint distribution is bimodal.

4.49 Behboodian (1990) illustrates how to construct bivariate random variables that are uncorrelated but dependent. Suppose that f_1, f_2, g_1, g_2 are univariate densities with means $\mu_1, \mu_2, \xi_1, \xi_2$, respectively, and the bivariate random variable (X, Y) has density

$$(X, Y) \sim af_1(x)g_1(y) + (1 - a)f_2(x)g_2(y),$$

where $0 < a < 1$ is known.

(a) Show that the marginal distributions are given by $f_X(x) = af_1(x) + (1 - a)f_2(x)$ and $f_Y(x) = ag_1(y) + (1 - a)g_2(y)$.

(b) Show that X and Y are independent if and only if $[f_1(x) - f_2(x)][g_1(y) - g_2(y)] = 0$.

(c) Show that $\text{Cov}(X, Y) = a(1-a)[\mu_1 - \mu_2][\xi_1 - \xi_2]$, and thus explain how to construct dependent uncorrelated random variables.

(d) Letting f_1, f_2, g_1, g_2 be binomial pmfs, give examples of combinations of parameters that lead to independent (X, Y) pairs, correlated (X, Y) pairs, and uncorrelated but dependent (X, Y) pairs.

4.50 If (X, Y) has the bivariate normal pdf

$$f(x,y) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right),$$

show that $\text{Corr}(X, Y) = \rho$ and $\text{Corr}(X^2, Y^2) = \rho^2$. (Conditional expectations will simplify calculations.)

4.51 Let X, Y, and Z be independent uniform$(0, 1)$ random variables.

(a) Find $P(X/Y \le t)$ and $P(XY \le t)$. (Pictures will help.)

(b) Find $P(XY/Z \le t)$.

4.52 Bullets are fired at the origin of an (x, y) coordinate system, and the point hit, say (X, Y), is a random variable. The variables X and Y are taken to be independent n$(0, 1)$ random variables. If two bullets are fired independently, what is the distribution of the distance between them?

4.53 Let A, B, and C be independent random variables, uniformly distributed on $(0, 1)$. What is the probability that Ax^2+Bx+C has real roots? (*Hint:* If $X \sim$ uniform$(0, 1)$, then $-\log X \sim$ exponential. The sum of two independent exponentials is gamma.)

4.54 Find the pdf of $\Pi_{i=1}^n X_i$, where the X_is are independent uniform$(0, 1)$ random variables. (*Hint:* Try to calculate the cdf, and remember the relationship between uniforms and exponentials.)

4.55 A *parallel system* is one that functions as long as at least one component of it functions. A particular parallel system is composed of three independent components, each of which has a lifelength with an exponential(λ) distribution. The lifetime of the system is the maximum of the individual lifelengths. What is the distribution of the lifetime of the system?

4.56 A large number, $N = mk$, of people are subject to a blood test. This can be administered in two ways.

(i) Each person can be tested separately. In this case N tests are required.

(ii) The blood samples of k people can be pooled and analyzed together. If the test is negative, this *one* test suffices for k people. If the test is positive, each of the k persons must be tested separately, and, in all, $k + 1$ tests are required for the k people.

Assume that the probability, p, that the test is positive is the same for all people and that the test results for different people are statistically independent.

(a) What is the probability that the test for a pooled sample of k people will be positive?

(b) Let $X =$ number of blood tests necessary under plan (ii). Find EX.

(c) In terms of minimizing the expected number of blood tests to be performed on the N people, which plan [(i) or (ii)] would be preferred if it is known that p is close to 0? Justify your answer using the expression derived in part (b).

4.57 Refer to Miscellanea 4.9.2.

(a) Show that A_1 is the arithmetic mean, A_{-1} is the harmonic mean, and $A_0 = \lim_{r\to 0} A_r$ is the geometric mean.

(b) The arithmetic-geometric-harmonic mean inequality will follow if it can be established that A_r is a nondecreasing function of r over the range $-\infty < r < \infty$.

(i) Verify that if $\log A_r$ is nondecreasing in r, then it will follow that A_r is nondecreasing in r.

(ii) Show that

$$\frac{d}{dr}\log A_r = \frac{1}{r^2}\left\{\frac{r\sum_i x_i^r \log x_i}{\sum_i x_i^r} - \log\left(\frac{1}{n}\sum_i x_i^r\right)\right\}.$$

(iii) Define $a_i = x_i^r/\sum_i x_i^r$ and write the quantity in braces as

$$\log(n) - \sum a_i \log(1/a_i),$$

where $\sum a_i = 1$. Now prove that this quantity is nonnegative, establishing the monotonicity of A_r and the arithmetic-geometric-harmonic mean inequality as a special case.

The quantity $\sum_i a_i \log(1/a_i)$ is called *entropy*, sometimes considered an absolute measure of uncertainty (see Bernardo and Smith 1994, Section 2.7). The result of part (iii) states that the maximum entropy is attained when all probabilities are the same (randomness).

(*Hint*: To prove the inequality note that the a_i are a probability distribution, and we can write

$$\mathrm{E}\log\left(\frac{1}{a}\right) = \sum_i a_i \log\left(\frac{1}{a_i}\right),$$

and Jensen's Inequality shows that $\mathrm{E}\log\left(\frac{1}{a}\right) \le \log\left(\mathrm{E}\frac{1}{a}\right)$.)

4.58 For any two random variables X and Y with finite variances, prove that

(a) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, \mathrm{E}(Y|X))$.

(b) X and $Y - \mathrm{E}(Y|X)$ are uncorrelated.

(c) $\mathrm{Var}(Y - \mathrm{E}(Y|X)) = \mathrm{E}(\mathrm{Var}(Y|X))$.

4.59 For any three random variables X, Y, and Z with finite variances, prove (in the sprit of Theorem 4.4.7) the covariance identity

$$\mathrm{Cov}(X,Y) = \mathrm{E}(\mathrm{Cov}(X,Y|Z)) + \mathrm{Cov}(\mathrm{E}(X|Z), \mathrm{E}(Y|Z)),$$

where $\mathrm{Cov}(X,Y|Z)$ is the covariance of X and Y under the pdf $f(x,y|z)$.

4.60 Referring to Miscellanea 4.9.3, find the conditional distribution of Y given that $Y = X$ for each of the three interpretations given for the condition $Y = X$.

4.61 DeGroot (1986) gives the following example of the Borel Paradox (Miscellanea 4.9.3): Suppose that X_1 and X_2 are iid exponential(1) random variables, and define $Z = (X_2 - 1)/X_1$. The probability-zero sets $\{Z = 0\}$ and $\{X_2 = 1\}$ seem to be giving us the same information but lead to different conditional distributions.

(a) Find the distribution of $X_1|Z = 0$, and compare it to the distribution of $X_1|X_2 = 1$.

(b) For small $\varepsilon > 0$ and $x_1 > 0, x_2 > 0$, consider the sets

$$B_1 = \{(x_1,x_2) : -\varepsilon < \frac{x_2 - 1}{x_1} < \varepsilon\} \quad \text{and} \quad B_2 = \{(x_1,x_2) : 1 - \varepsilon < x_2 < 1 + \varepsilon\}.$$

Draw these sets and support the argument that B_1 is informative about X_1 but B_2 is not.

(c) Calculate $P(X_1 \le x|B_1)$ and $P(X_1 \le x|B_2)$, and show that their limits (as $\varepsilon \to 0$) agree with part (a).

(*Communicated by L. Mark Berliner, Ohio State University.*)

4.62 Finish the proof of the equality in Jensen's Inequality (Theorem 4.7.7). Let $g(x)$ be a convex function. Suppose $a+bx$ is a line tangent to $g(x)$ at $x = \mathrm{E}\,X$, and $g(x) > a+bx$ except at $x = \mathrm{E}\,X$. Then $\mathrm{E}\,g(X) > g(\mathrm{E}\,X)$ unless $P(X = \mathrm{E}\,X) = 1$.

4.63 A random variable X is defined by $Z = \log X$, where $\mathrm{E}Z = 0$. Is $\mathrm{E}X$ greater than, less than, or equal to 1?

4.64 This exercise involves a well-known inequality known as the *triangle inequality* (a special case of Minkowski's Inequality).

(a) Prove (without using Minkowski's Inequality) that for any numbers a and b

$$|a+b| \le |a| + |b|.$$

(b) Use part (a) to establish that for any random variables X and Y with finite expectations,

$$\mathrm{E}|X+Y| \le \mathrm{E}|X| + \mathrm{E}|Y|.$$

4.65 Prove the Covariance Inequality by generalizing the argument given in the text immediately preceding the inequality.

4.9 Miscellanea

4.9.1 The Exchange Paradox

The *"Exchange Paradox"* (Christensen and Utts 1992) has generated a lengthy dialog among statisticians. The problem (or the paradox) goes as follows:

> A swami puts m dollars in one envelope and $2m$ dollars in another. You and your opponent each get one of the envelopes (at random). You open your envelope and find x dollars, and then the swami asks you if you want to trade envelopes. You reason that if you switch, you will get either $x/2$ or $2x$ dollars, each with probability $1/2$. This makes the expected value of a switch equal to $(1/2)(x/2) + (1/2)(2x) = 5x/4$, which is greater than the x dollars that you hold in your hand. So you offer to trade.
>
> The paradox is that your opponent has done the same calculation. How can the trade be advantageous for both of you?

(i) Christensen and Utts say, "The conclusion that trading envelopes is always optimal is based on the assumption that there is no information obtained by observing the contents of the envelope," and they offer the following resolution. Let $M \sim \pi(m)$ be the pdf for the amount of money placed in the first envelope, and let X be the amount of money in your envelope. Then $P(X = m|M = m) = P(X = 2m|M = m) = 1/2$, and hence

$$P(M = x|X = x) = \frac{\pi(x)}{\pi(x) + \pi(x/2)} \text{ and } P(M = x/2|X = x) = \frac{\pi(x/2)}{\pi(x) + \pi(x/2)}.$$

It then follows that the expected winning from a trade is

$$\frac{\pi(x)}{\pi(x)+\pi(x/2)}2x + \frac{\pi(x/2)}{\pi(x)+\pi(x/2)}\frac{x}{2},$$

and thus you should trade only if $\pi(x/2) < 2\pi(x)$. If π is the exponential(λ) density, it is optimal to trade if $x < 2\log 2/\lambda$.

(ii) A more classical approach does not assume that there is a pdf on the amount of money placed in the first envelope. Christensen and Utts also offer an explanation here, noting that the paradox occurs if one incorrectly assumes that $P(Y = y|X = x) = 1/2$ for all values of X and Y, where X is the amount in your envelope and Y is the amount in your opponent's envelope. They argue that the correct conditional distributions are $P(Y = 2x|X = m) = 1$ and $P(Y = x/2|X = 2m) = 1$ and that your expected winning if you trade is $E(Y) = 3m/2$, which is the same as your expected winning if you keep your envelope.

This paradox is often accompanied with arguments for or against the Bayesian methodology of inference (see Chapter 7), but these arguments are somewhat tangential to the underlying probability calculations. For comments, criticisms, and other analyses see the letters to the editor from Binder (1993), Ridgeway (1993) (which contains a solution by Marilyn vos Savant), Ross (1994), and Blachman (1996) and the accompanying responses from Christensen and Utts.

4.9.2 More on the Arithmetic-Geometric-Harmonic Mean Inequality

The arithmetic-geometric-harmonic mean inequality is a special case of a general result about *power means*, which are defined by

$$A_r = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^r\right]^{1/r}$$

for $x_i \geq 0$. Shier (1988) shows that A_r is a nondecreasing function of r; that is, $A_r \leq A_{r'}$ if $r \leq r'$ or

$$\left[\frac{1}{n}\sum_{i=1}^{n} x_i^r\right]^{1/r} \leq \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{r'}\right]^{1/r'} \qquad \text{for } r \leq r'.$$

It should be clear that A_1 is the arithmetic mean and A_{-1} is the harmonic mean. What is less clear, but true, is that $A_0 = \lim_{r\to 0} A_r$ is the geometric mean. Thus, the arithmetic-geometric-harmonic mean inequality follows as a special case of the power mean inequality (see Exercise 4.57).

4.9.3 The Borel Paradox

Throughout this chapter, for continuous random variables X and Y, we have been writing expressions such as $\mathrm{E}(Y|X = x)$ and $P(Y \leq y|X = x)$. Thus far, we have not gotten into trouble. However, we might have.

Formally, the conditioning in a conditional expectation is done with respect to a sub sigma-algebra (Definition 1.2.1), and the conditional expectation $\mathrm{E}(Y|\mathcal{G})$ is

defined as a random variable whose integral, over any set in the sub sigma-algebra $\mathcal{G}$, agrees with that of X. This is quite an advanced concept in probability theory (see Billingsley 1995, Section 34).

Since the conditional expectation is only defined in terms of its integral, it may not be unique even if the conditioning is well-defined. However, when we condition on sets of probability 0 (such as $\{X = x\}$), conditioning may not be well defined, so different conditional expectations are more likely to appear. To see how this could affect us, it is easiest to look at conditional distributions, which amounts to calculating $\mathrm{E}[I(Y \leq y)|X = x]$.

Proschan and Presnell (1998) tell the story of a statistics exam that had the question "If X and Y are independent standard normals, what is the conditional distribution of Y given that $Y = X$?" Different students interpreted the condition $Y = X$ in the following ways:

(1) $Z_1 = 0$, where $Z_1 = Y - X$;
(2) $Z_2 = 1$, where $Z_2 = Y/X$;
(3) $Z_3 = 1$, where $Z_3 = I(Y = X)$.

Each condition is a correct interpretation of the condition $Y = X$, and each leads to a different conditional distribution (see Exercise 4.60).

This is the *Borel Paradox* and arises because different (correct) interpretations of the probability 0 conditioning sets result in different conditional expectations. How can we avoid the paradox? One way is to avoid conditioning on sets of probability 0. That is, compute only $\mathrm{E}(Y|X \in B)$, where B is a set with $P(X \in B) > 0$. So to compute something like $\mathrm{E}(Y|X = x)$, take a sequence $B_n \downarrow x$, and define $\mathrm{E}(Y|X = x) = \lim_{n\to\infty} \mathrm{E}(Y|X \in B_n)$. We now avoid the paradox, as the different answers for $\mathrm{E}(Y|X = x)$ will arise from different sequences, so there should be no surprises (Exercise 4.61).