Chapter 3
Expectation

CHAPTER OUTLINE

In the first two chapters we learned about probability models, random variables, and distributions. There is one more concept that is fundamental to all of probability theory, that of expected value.

Intuitively, the expected value of a random variable is the average value that the random variable takes on. For example, if half the time $X = 0$, and the other half of the time $X = 10$, then the average value of X is 5. We shall write this as $E(X) = 5$. Similarly, if one-third of the time $Y = 6$ while two-thirds of the time $Y = 15$, then $E(Y) = 12$.

Another interpretation of expected value is in terms of fair gambling. Suppose someone offers you a ticket (e.g., a lottery ticket) worth a certain random amount X. How much would you be willing to pay to buy the ticket? It seems reasonable that you would be willing to pay the expected value $E(X)$ of the ticket, but no more. However, this interpretation does have certain limitations; see Example 3.1.12.

To understand expected value more precisely, we consider discrete and absolutely continuous random variables separately.

3.1 The Discrete Case

We begin with a definition.

Definition 3.1.1 Let X be a discrete random variable. Then the *expected value* (or *mean value* or *mean*) of X, written $E(X)$ (or μ_X), is defined by

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

We will have $P(X = x) = 0$ except for those values x that are possible values of X. Hence, an equivalent definition is the following.

Definition 3.1.2 Let X be a discrete random variable, taking on distinct values $x_1, x_2, \ldots$, with $p_i = P(X = x_i)$. Then the *expected value* of X is given by

$$E(X) = \sum_i x_i\, p_i.$$

The definition (in either form) is best understood through examples.

EXAMPLE 3.1.1
Suppose, as above, that $P(X = 0) = P(X = 10) = 1/2$. Then

$$E(X) = (0)(1/2) + (10)(1/2) = 5,$$

as predicted. ∎

EXAMPLE 3.1.2
Suppose, as above, that $P(Y = 6) = 1/3$, and $P(Y = 15) = 2/3$. Then

$$E(Y) = (6)(1/3) + (15)(2/3) = 2 + 10 = 12,$$

again as predicted. ∎

EXAMPLE 3.1.3
Suppose that $P(Z = -3) = 0.2$, and $P(Z = 11) = 0.7$, and $P(Z = 31) = 0.1$. Then

$$E(Z) = (-3)(0.2) + (11)(0.7) + (31)(0.1) = -0.6 + 7.7 + 3.1 = 10.2.$$ ∎

EXAMPLE 3.1.4
Suppose that $P(W = -3) = 0.2$, and $P(W = -11) = 0.7$, and $P(W = 31) = 0.1$. Then

$$E(W) = (-3)(0.2) + (-11)(0.7) + (31)(0.1) = -0.6 - 7.7 + 3.1 = -5.2.$$

In this case, the expected value of W is *negative*. ∎

We thus see that, for a discrete random variable X, once we know the probabilities that $X = x$ (or equivalently, once we know the probability function p_X), it is straightforward (at least in simple cases) to compute the expected value of X.

We now consider some of the common discrete distributions introduced in Section 2.3.

EXAMPLE 3.1.5 *Degenerate Distributions*
If $X \equiv c$ is a constant, then $P(X = c) = 1$, so

$$E(X) = (c)(1) = c,$$

as it should. ∎

EXAMPLE 3.1.6 *The Bernoulli(θ) Distribution and Indicator Functions*
If $X \sim$ Bernoulli(θ), then $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$, so

$$E(X) = (1)(\theta) + (0)(1 - \theta) = \theta.$$

As a particular application of this, suppose we have a response s taking values in a sample S and $A \subset S$. Letting $X(s) = I_A(s)$, we have that X is the indicator function of the set A and so takes the values 0 and 1. Then we have that $P(X = 1) = P(A)$, and so $X \sim$ Bernoulli($P(A)$). This implies that

$$E(X) = E(I_A) = P(A).$$

Therefore, we have shown that the expectation of the indicator function of the set A is equal to the probability of A. ∎

EXAMPLE 3.1.7 *The Binomial(n, θ) Distribution*
If $Y \sim$ Binomial(n, θ), then

$$P(Y = k) = \binom{n}{k}\theta^k(1 - \theta)^{n-k}$$

for $k = 0, 1, \ldots, n$. Hence,

$$\begin{aligned} E(Y) &= \sum_{k=0}^{n} k\, P(Y = k) = \sum_{k=0}^{n} k\binom{n}{k}\theta^k(1 - \theta)^{n-k} \\ &= \sum_{k=0}^{n} k\,\frac{n!}{k!\,(n-k)!}\theta^k(1 - \theta)^{n-k} = \sum_{k=1}^{n}\frac{n!}{(k-1)!\,(n-k)!}\theta^k(1 - \theta)^{n-k} \\ &= \sum_{k=1}^{n}\frac{n\,(n-1)!}{(k-1)!\,(n-k)!}\theta^k(1 - \theta)^{n-k} = \sum_{k=1}^{n} n\binom{n-1}{k-1}\theta^k(1 - \theta)^{n-k}. \end{aligned}$$

Now, the *binomial theorem* says that for any a and b and any positive integer m,

$$(a + b)^m = \sum_{j=0}^{m}\binom{m}{j}a^j b^{m-j}.$$

Using this, and setting $j = k - 1$, we see that

$$\begin{aligned} E(Y) &= \sum_{k=1}^{n} n\binom{n-1}{k-1}\theta^k(1 - \theta)^{n-k} = \sum_{j=0}^{n-1} n\binom{n-1}{j}\theta^{j+1}(1 - \theta)^{n-j-1} \\ &= n\theta\sum_{j=0}^{n-1}\binom{n-1}{j}\theta^j(1 - \theta)^{n-j-1} = n\theta\left((\theta + 1 - \theta)^{n-1}\right) = n\theta. \end{aligned}$$

Hence, the expected value of Y is $n\theta$. Note that this is precisely n times the expected value of X, where $X \sim \text{Bernoulli}(\theta)$ as in Example 3.1.6. We shall see in Example 3.1.15 that this is not a coincidence. ∎

EXAMPLE 3.1.8 *The Geometric(θ) Distribution*
If $Z \sim \text{Geometric}(\theta)$, then $P(Z = k) = (1-\theta)^k\theta$ for $k = 0, 1, 2, \ldots$. Hence,

$$E(Z) = \sum_{k=0}^{\infty} k(1-\theta)^k\theta. \tag{3.1.1}$$

Therefore, we can write

$$(1-\theta)E(Z) = \sum_{\ell=0}^{\infty} \ell(1-\theta)^{\ell+1}\theta.$$

Using the substitution $k = \ell + 1$, we compute that

$$(1-\theta)E(Z) = \sum_{k=1}^{\infty} (k-1)\,(1-\theta)^k\theta. \tag{3.1.2}$$

Subtracting (3.1.2) from (3.1.1), we see that

$$\begin{aligned} \theta E(Z) &= (E(Z)) - ((1-\theta)E(Z)) = \sum_{k=1}^{\infty} (k-(k-1))\,(1-\theta)^k\theta \\ &= \sum_{k=1}^{\infty} (1-\theta)^k\theta = \frac{1-\theta}{1-(1-\theta)}\theta = 1-\theta. \end{aligned}$$

Hence, $\theta E(Z) = 1-\theta$, and we obtain $E(Z) = (1-\theta)/\theta$. ∎

EXAMPLE 3.1.9 *The Poisson(λ) Distribution*
If $X \sim \text{Poisson}(\lambda)$, then $P(X = k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, 2, \ldots$. Hence, setting $\ell = k - 1$,

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda\, e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{\lambda^\ell}{\ell!} = \lambda\, e^{-\lambda}e^{\lambda} = \lambda, \end{aligned}$$

and we conclude that $E(X) = \lambda$. ∎

It should be noted that expected values can sometimes be infinite, as the following example demonstrates.

EXAMPLE 3.1.10
Let X be a discrete random variable, with probability function p_X given by

$$p_X(2^k) = 2^{-k}$$

for $k = 1, 2, 3, \ldots$, with $p_X(x) = 0$ for other values of x. That is, $p_X(2) = 1/2$, $p_X(4) = 1/4$, $p_X(8) = 1/8$, etc., while $p_X(1) = p_X(3) = p_X(5) = p_X(6) = \cdots = 0$.

Then it is easily checked that p_X is indeed a valid probability function (i.e., $p_X(x) \geq 0$ for all x, with $\sum_x p_X(x) = 1$). On the other hand, we compute that

$$E(X) = \sum_{k=1}^{\infty} (2^k)(2^{-k}) = \sum_{k=1}^{\infty} (1) = \infty.$$

We therefore say that $E(X) = \infty$, i.e., that the expected value of X is infinite. ∎

Sometimes the expected value simply does not exist, as in the following example.

EXAMPLE 3.1.11
Let Y be a discrete random variable, with probability function p_Y given by

$$p_Y(y) = \begin{cases} 1/2y & y = 2, 4, 8, 16, \ldots \\ 1/2|y| & y = -2, -4, -8, -16, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

That is, $p_Y(2) = p_Y(-2) = 1/4$, $p_Y(4) = p_Y(-4) = 1/8$, $p_Y(8) = p_Y(-8) = 1/16$, etc. Then it is easily checked that p_Y is indeed a valid probability function (i.e., $p_Y(y) \geq 0$ for all y, with $\sum_y p_Y(y) = 1$).

On the other hand, we compute that

$$\begin{aligned} E(Y) &= \sum_y y\, p_Y(y) = \sum_{k=1}^{\infty} (2^k)(1/2\,2^k) + \sum_{k=1}^{\infty} (-2^k)(1/2\,2^k) \\ &= \sum_{k=1}^{\infty} (1/2) - \sum_{k=1}^{\infty} (1/2) = \infty - \infty, \end{aligned}$$

which is *undefined*. We therefore say that $E(Y)$ does not exist, i.e., that the expected value of Y is *undefined* in this case. ∎

EXAMPLE 3.1.12 *The St. Petersburg Paradox*
Suppose someone makes you the following deal. You will repeatedly flip a fair coin and will receive an award of 2^Z pennies, where Z is the number of tails that appear before the first head. How much would you be willing to pay for this deal?

Well, the probability that the award will be 2^z pennies is equal to the probability that you will flip z tails and then one head, which is equal to $1/2^{z+1}$. Hence, the expected value of the award (in pennies) is equal to

$$\sum_{z=0}^{\infty} (2^z)(1/2^{z+1}) = \sum_{z=0}^{\infty} 1/2 = \infty.$$

In words, the average amount of the award is infinite!

Hence, according to the "fair gambling" interpretation of expected value, as discussed at the beginning of this chapter, it seems that you should be willing to pay an infinite amount (or, at least, any finite amount no matter how large) to get the award

promised by this deal! How much do you think you should *really* be willing to pay for it?[1] ∎

EXAMPLE 3.1.13 *The St. Petersburg Paradox, Truncated*
Suppose in the St. Petersburg paradox (Example 3.1.12), it is agreed that the award will be truncated at 2^{30} cents (which is just over $10 million!). That is, the award will be the same as for the original deal, except the award will be frozen once it exceeds 2^{30} cents. Formally, the award is now equal to $2^{\min(30,Z)}$ pennies, where Z is as before.

How much would you be willing to pay for this new award? Well, the expected value of the new award (in cents) is equal to

$$\begin{aligned}\sum_{z=1}^{\infty}(2^{\min(30,z)})(1/2^{z+1}) &= \sum_{z=1}^{30}(2^z)(1/2^{z+1}) + \sum_{z=31}^{\infty}(2^{30})(1/2^{z+1}) \\ &= \sum_{z=1}^{30}(1/2) + (2^{30})(1/2^{31}) = 31/2 = 15.5.\end{aligned}$$

That is, truncating the award at just over $10 million changes its expected value enormously, from infinity to less than 16 cents! ∎

In *utility theory*, it is often assumed that each person has a utility function U such that, if they win x cents, their amount of "utility" (i.e., benefit or joy or pleasure) is equal to $U(x)$. In this context, the truncation of Example 3.1.13 may be thought of not as changing the rules of the game but as corresponding to a utility function of the form $U(x) = \min(x, 2^{30})$. In words, this says that your utility is equal to the amount of money you get, until you reach 2^{30} cents (approximately $10 million), after which point you don't care about money[2] anymore. The result of Example 3.1.13 then says that, with this utility function, the St. Petersburg paradox is only worth 15.5 cents to you — even though its expected value is infinite.

We often need to compute expected values of *functions* of random variables. Fortunately, this is not too difficult, as the following theorem shows.

Theorem 3.1.1
(a) Let X be a discrete random variable, and let $g : R^1 \to R^1$ be some function such that the expectation of the random variable $g(X)$ exists. Then

$$E\,(g(X)) = \sum_x g(x)\,P(X = x).$$

(b) Let X and Y be discrete random variables, and let $h : R^2 \to R^1$ be some function such that the expectation of the random variable $h(X, Y)$ exists. Then

$$E\,(h(X, Y)) = \sum_{x,y} h(x, y)P(X = x,\ Y = y).$$

[1] When one of the authors first heard about this deal, he decided to try it and agreed to pay $1. In fact, he got four tails before the first head, so his award was 16 cents, but he still lost 84 cents overall.

[2] Or, perhaps, you think it is unlikely you will be able to *collect* the money!

PROOF We prove part (b) here. Part (a) then follows by simply setting $h(x, y) = g(x)$ and noting that

$$\sum_{x,y} g(x)\, P(X = x,\ Y = y) = \sum_{x} g(x)\, P(X = x).$$

Let $Z = h(X, Y)$. We have that

$$\begin{aligned} E(Z) &= \sum_{z} z\, P(Z = z) = \sum_{z} z\, P(h(X, Y) = z) \\ &= \sum_{z} z \sum_{\substack{x,y \\ h(x,y)=z}} P(X = x,\ Y = y) = \sum_{x,y} \sum_{\substack{z \\ z=h(x,y)}} z\, P(X = x,\ Y = y) \\ &= \sum_{x,y} h(x, y)\, P(X = x,\ Y = y), \end{aligned}$$

as claimed. ∎

One of the most important properties of expected value is that it is *linear*, stated as follows.

Theorem 3.1.2 (*Linearity of expected values*) Let X and Y be discrete random variables, let a and b be real numbers, and put $Z = aX + bY$. Then $E(Z) = aE(X) + bE(Y)$.

PROOF Let $p_{X,Y}$ be the joint probability function of X and Y. Then using Theorem 3.1.1,

$$\begin{aligned} E(Z) &= \sum_{x,y} (ax + by)\, p_{X,Y}(x, y) = a \sum_{x,y} x\, p_{X,Y}(x, y) + b \sum_{x,y} y\, p_{X,Y}(x, y) \\ &= a \sum_{x} x \sum_{y} p_{X,Y}(x, y) + b \sum_{y} y \sum_{x} p_{X,Y}(x, y). \end{aligned}$$

Because $\sum_y p_{X,Y}(x, y) = p_X(x)$ and $\sum_x p_{X,Y}(x, y) = p_Y(y)$, we have that

$$E(Z) = a \sum_{x} x\, p_X(x) + b \sum_{y} y\, p_Y(y) = aE(X) + bE(Y),$$

as claimed. ∎

EXAMPLE 3.1.14
Let $X \sim \text{Binomial}(n, \theta_1)$, and let $Y \sim \text{Geometric}(\theta_2)$. What is $E(3X - 2Y)$?

We already know (Examples 3.1.6 and 3.1.7) that $E(X) = n\,\theta_1$ and $E(Y) = (1 - \theta_2)\,/\,\theta_2$. Hence, by Theorem 3.1.2, $E(3X - 2Y) = 3E(X) - 2E(Y) = 3n\theta_1 - 2(1 - \theta_2)/\theta_2$.∎

EXAMPLE 3.1.15
Let $Y \sim \text{Binomial}(n, \theta)$. Then we know (cf. Example 2.3.3) that we can think of $Y = X_1 + \cdots + X_n$, where each $X_i \sim \text{Bernoulli}(\theta)$ (in fact, $X_i = 1$ if the ith coin is

heads, otherwise $X_i = 0$). Because $E(X_i) = \theta$ for each i, it follows immediately from Theorem 3.1.2 that

$$E(Y) = E(X_1) + \cdots + E(X_n) = \theta + \cdots + \theta = n\theta.$$

This gives the same answer as Example 3.1.7, but much more easily. ∎

Suppose that X is a random variable and $Y = c$ is a constant. Then from Theorem 3.1.2, we have that $E(X + c) = E(X) + c$. From this we see that the mean value μ_X of X is a measure of the *location* of the probability distribution of X. For example, if X takes the value x with probability p and the value y with probability $1 - p$, then the mean of X is $\mu_X = px + (1 - p)y$ which is a value between x and y. For a constant c, the probability distribution of $X + c$ is concentrated on the points $x + c$ and $y + c$, with probabilities p and $1 - p$, respectively. The mean of $X + c$ is $\mu_X + c$, which is between the points $x + c$ and $y + c$, i.e., the mean shifts with the probability distribution. It is also true that if X is concentrated on the finite set of points $x_1 < x_2 < \cdots < x_k$, then $x_1 \leq \mu_X \leq x_k$, and the mean shifts exactly as we shift the distribution. This is depicted in Figure 3.1.1 for a distribution concentrated on $k = 4$ points. Using the results of Section 2.6.1, we have that $p_{X+c}(x) = p_X(x - c)$.

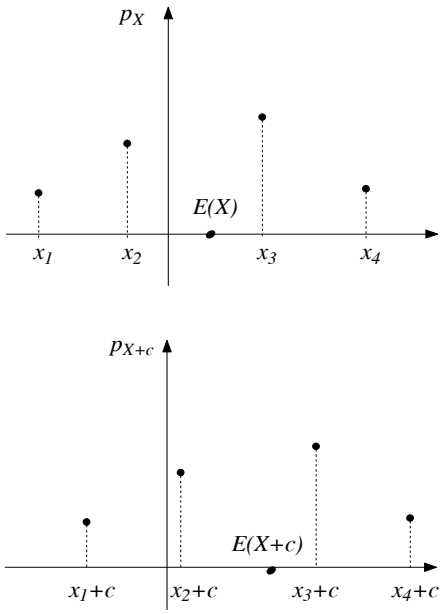

Figure 3.1.1: The probability functions and means of discrete random variables X and $X + c$.

Theorem 3.1.2 says, in particular, that $E(X + Y) = E(X) + E(Y)$, i.e., that expectation preserves sums. It is reasonable to ask whether the same property holds for products. That is, do we necessarily have $E(XY) = E(X)E(Y)$? In general, the answer is no, as the following example shows.

EXAMPLE 3.1.16
Let X and Y be discrete random variables, with joint probability function given by

$$p_{X,Y}(x, y) = \begin{cases} 1/2 & x = 3,\ y = 5 \\ 1/6 & x = 3,\ y = 9 \\ 1/6 & x = 6,\ y = 5 \\ 1/6 & x = 6,\ y = 9 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E(X) = \sum_x x P(X = x) = (3)(1/2 + 1/6) + (6)(1/6 + 1/6) = 4$$

and

$$E(Y) = \sum_y y\, P(Y = y) = (5)(1/2 + 1/6) + (9)(1/6 + 1/6) = 19/3,$$

while

$$\begin{aligned} E(XY) &= \sum_z z\, P(XY = z) \\ &= (3)(5)(1/2) + (3)(9)(1/6) + (6)(5)(1/6) + (6)(9)(1/6) \\ &= 26. \end{aligned}$$

Because $(4)(19/3) \neq 26$, we see that $E(X)\, E(Y) \neq E(XY)$ in this case. ∎

On the other hand, if X and Y are *independent*, then we do have $E(X)E(Y) = E(XY)$.

Theorem 3.1.3 Let X and Y be discrete random variables that are independent. Then $E(XY) = E(X)E(Y)$.

PROOF Independence implies (see Theorem 2.8.3) that $P(X = x,\ Y = y) = P(X = x)\, P(Y = y)$. Using this, we compute by Theorem 3.1.1 that

$$\begin{aligned} E(XY) &= \sum_{x,y} xy\, P(X = x,\ Y = y) = \sum_{x,y} xy\, P(X = x)\, P(Y = y) \\ &= \left(\sum_x x\, P(X = x) \right) \left(\sum_y y\, P(Y = y) \right) = E(X)\, E(Y), \end{aligned}$$

as claimed. ∎

Theorem 3.1.3 will be used often in subsequent chapters, as will the following important property.

Theorem 3.1.4 (*Monotonicity*) Let X and Y be discrete random variables, and suppose that $X \leq Y$. (Remember that this means $X(s) \leq Y(s)$ for all $s \in S$.) Then $E(X) \leq E(Y)$.

PROOF Let $Z = Y - X$. Then Z is also discrete. Furthermore, because $X \leq Y$, we have $Z \geq 0$, so that all possible values of Z are nonnegative. Hence, if we list the possible values of Z as $z_1, z_2, \ldots$, then $z_i \geq 0$ for all i, so that

$$E(Z) = \sum_i z_i P(Z = z_i) \geq 0.$$

But by Theorem 3.1.2, $E(Z) = E(Y) - E(X)$. Hence, $E(Y) - E(X) \geq 0$, so that $E(Y) \geq E(X)$. ∎

Summary of Section 3.1

- The expected value $E(X)$ of a random variable X represents the long-run average value that it takes on.
- If X is discrete, then $E(X) = \sum_x x\, P(X = x)$.
- The expected values of the Bernoulli, binomial, geometric, and Poisson distributions were computed.
- Expected value has an interpretation in terms of fair gambling, but such interpretations require utility theory to accurately reflect human behavior.
- Expected values of functions of one or two random variables can also be computed by summing the function values times the probabilities.
- Expectation is linear and monotone.
- If X and Y are independent, then $E(XY) = E(X)\,E(Y)$. But without independence, this property may fail.

EXERCISES

3.1.1 Compute $E(X)$ when the probability function of X is given by each of the following.

(a)

$$p_X(x) = \begin{cases} 1/7 & x = -4 \\ 2/7 & x = 0 \\ 4/7 & x = 3 \\ 0 & \text{otherwise} \end{cases}$$

(b)

$$p_X(x) = \begin{cases} 2^{-x-1} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

(c)

$$p_X(x) = \begin{cases} 2^{-x+6} & x = 7, 8, 9, \ldots \\ 0 & \text{otherwise} \end{cases}$$

3.1.2 Let X and Y have joint probability function given by

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

as in Example 2.7.5. Compute each of the following.
(a) $E(X)$
(b) $E(Y)$
(c) $E(3X+7Y)$
(d) $E(X^2)$
(e) $E(Y^2)$
(f) $E(XY)$
(g) $E(XY+14)$

3.1.3 Let X and Y have joint probability function given by

$$p_{X,Y}(x,y) = \begin{cases} 1/2 & x=2,\ y=10 \\ 1/6 & x=-7,\ y=10 \\ 1/12 & x=2,\ y=12 \\ 1/12 & x=-7,\ y=12 \\ 1/12 & x=2,\ y=14 \\ 1/12 & x=-7,\ y=14 \\ 0 & \text{otherwise.} \end{cases}$$

Compute each of the following.
(a) $E(X)$.
(b) $E(Y)$
(c) $E(X^2)$
(d) $E(Y^2)$
(e) $E(X^2+Y^2)$
(f) $E(XY-4Y)$

3.1.4 Let $X \sim \text{Bernoulli}(\theta_1)$ and $Y \sim \text{Binomial}(n,\theta_2)$. Compute $E(4X-3Y)$.

3.1.5 Let $X \sim \text{Geometric}(\theta)$ and $Y \sim \text{Poisson}(\lambda)$. Compute $E(8X-Y+12)$.

3.1.6 Let $Y \sim \text{Binomial}(100,\ 0.3)$, and $Z \sim \text{Poisson}(7)$. Compute $E(Y+Z)$.

3.1.7 Let $X \sim \text{Binomial}(80,\ 1/4)$, and let $Y \sim \text{Poisson}(3/2)$. Assume X and Y are independent. Compute $E(XY)$.

3.1.8 Starting with one penny, suppose you roll one fair six-sided die and get paid an additional number of pennies equal to three times the number showing on the die. Let X be the total number of pennies you have at the end. Compute $E(X)$.

3.1.9 Suppose you start with eight pennies and flip one fair coin. If the coin comes up heads, you get to keep all your pennies; if the coin comes up tails, you have to give half of them back. Let X be the total number of pennies you have at the end. Compute $E(X)$.

3.1.10 Suppose you flip two fair coins. Let $Y = 3$ if the two coins show the same result, otherwise let $Y = 5$. Compute $E(Y)$.

3.1.11 Suppose you roll two fair six-sided dice.

(a) Let Z be the sum of the two numbers showing. Compute $E(Z)$.

(b) Let W be the product of the two numbers showing. Compute $E(W)$.

3.1.12 Suppose you flip one fair coin and roll one fair six-sided die. Let X be the product of the numbers of heads (i.e., 0 or 1) times the number showing on the die. Compute $E(X)$. (Hint: Do not forget Theorem 3.1.3.)

3.1.13 Suppose you roll one fair six-sided die and then flip as many coins as the number showing on the die. (For example, if the die shows 4, then you flip four coins.) Let Y be the number of heads obtained. Compute $E(Y)$.

3.1.14 Suppose you roll three fair coins, and let X be the *cube* of the number of heads showing. Compute $E(X)$.

PROBLEMS

3.1.15 Suppose you start with one penny and repeatedly flip a fair coin. Each time you get heads, *before* the first time you get tails, you get two more pennies. Let X be the total number of pennies you have at the end. Compute $E(X)$.

3.1.16 Suppose you start with one penny and repeatedly flip a fair coin. Each time you get heads, *before* the first time you get tails, your number of pennies is *doubled*. Let X be the total number of pennies you have at the end. Compute $E(X)$.

3.1.17 Let $X \sim \text{Geometric}(\theta)$, and let $Y = \min(X, 100)$.
(a) Compute $E(Y)$.
(b) Compute $E(Y - X)$.

3.1.18 Give an example of a random variable X such that $E\big(\min(X, 100)\big) = E(X)$.

3.1.19 Give an example of a random variable X such that $E(\min(X, 100)) = E(X)/2$.

3.1.20 Give an example of a joint probability function $p_{X,Y}$ for random variables X and Y, such that $X \sim \text{Bernoulli}(1/4)$ and $Y \sim \text{Bernoulli}(1/2)$, but $E(XY) \neq 1/8$.

3.1.21 For $X \sim \text{Hypergeometric}(N, M, n)$, prove that $E(X) = nM/N$.

3.1.22 For $X \sim \text{Negative-Binomial}(r, \theta)$, prove that $E(X) = r(1-\theta)/\theta$. (Hint: Argue that if $X_1, \ldots, X_r$ are independent and identically distributed Geometric(θ), then $X = X_1 + \cdots + X_r \sim \text{Negative-Binomial}(r, \theta)$.)

3.1.23 Suppose that $(X_1, X_2, X_3) \sim \text{Multinomial}(n, \theta_1, \theta_2, \theta_3)$. Prove that $E(X_i) = n\theta_i$.

CHALLENGES

3.1.24 Let $X \sim \text{Geometric}(\theta)$. Compute $E(X^2)$.

3.1.25 Suppose X is a discrete random variable, such that $E(\min(X, M)) = E(X)$. Prove that $P(X > M) = 0$.

DISCUSSION TOPICS

3.1.26 How much would *you* be willing to pay for the deal corresponding to the St. Petersburg paradox (see Example 3.1.12)? Justify your answer.

3.1.27 What utility function U (as in the text following Example 3.1.13) best describes your own personal attitude toward money? Why?

3.2 The Absolutely Continuous Case

Suppose now that X is absolutely continuous, with density function f_X. How can we compute $E(X)$ then? By analogy with the discrete case, we might try computing $\sum_x x\, P(X = x)$, but because $P(X = x)$ is always zero, this sum is always zero as well.

On the other hand, if ϵ is a small positive number, then we could try approximating $E(X)$ by

$$E(X) \approx \sum_i i\epsilon\, P(i\epsilon \le X < (i+1)\epsilon),$$

where the sum is over all integers i. This makes sense because, if ϵ is small and $i\epsilon \le X < (i+1)\epsilon$, then $X \approx i\epsilon$.

Now, we know that

$$P(i\epsilon \le X < (i+1)\epsilon) = \int_{i\epsilon}^{(i+1)\epsilon} f_X(x)\, dx.$$

This tells us that

$$E(X) \approx \sum_i \int_{i\epsilon}^{(i+1)\epsilon} i\epsilon\, f_X(x)\, dx.$$

Furthermore, in this integral, $i\epsilon \le x < (i+1)\epsilon$. Hence, $i\epsilon \approx x$. We therefore see that

$$E(X) \approx \sum_i \int_{i\epsilon}^{(i+1)\epsilon} x\, f_X(x)\, dx = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$$

This prompts the following definition.

Definition 3.2.1 Let X be an absolutely continuous random variable, with density function f_X. Then the *expected value* of X is given by

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$$

From this definition, it is not too difficult to compute the expected values of many of the standard absolutely continuous distributions.

EXAMPLE 3.2.1 *The Uniform*[0, 1] *Distribution*
Let $X \sim$ Uniform[0, 1] so that the density of X is given by

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

Hence,

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_0^1 x\, dx = \frac{x^2}{2}\Big|_{x=0}^{x=1} = 1/2,$$

as one would expect. ∎

EXAMPLE 3.2.2 *The Uniform*$[L, R]$ *Distribution*
Let $X \sim$ Uniform$[L, R]$ so that the density of X is given by

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

Hence,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_L^R x\, \frac{1}{R-L}\, dx = \frac{x^2}{2(R-L)}\Big|_{x=L}^{x=R} \\ &= \frac{R^2 - L^2}{2(R-L)} = \frac{(R-L)(R+L)}{2(R-L)} = \frac{R+L}{2}, \end{aligned}$$

again as one would expect. ∎

EXAMPLE 3.2.3 *The Exponential*(λ) *Distribution*
Let $Y \sim$ Exponential(λ) so that the density of Y is given by

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & y \ge 0 \\ 0 & y < 0. \end{cases}$$

Hence, integration by parts, with $u = y$ and $dv = \lambda\, e^{-\lambda y}$ (so $du = dy$, $v = -e^{-\lambda y}$), leads to

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \int_0^{\infty} y\, \lambda\, e^{-\lambda y}\, dy = -\, y e^{-\lambda y}\big|_0^{\infty} + \int_0^{\infty} e^{-\lambda y}\, dy \\ &= \int_0^{\infty} e^{-\lambda y}\, dy = -\, \frac{e^{-\lambda y}}{\lambda}\Big|_0^{\infty} = -\frac{0-1}{\lambda} = \frac{1}{\lambda}. \end{aligned}$$

In particular, if $\lambda = 1$, then $Y \sim$ Exponential(1), and $E(Y) = 1$. ∎

EXAMPLE 3.2.4 *The* $N(0, 1)$ *Distribution*
Let $Z \sim N(0, 1)$ so that the density of Z is given by

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

Hence,

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z\, f_Z(z)\, dz \\ &= \int_{-\infty}^{\infty} z\, \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\ &= \int_{-\infty}^{0} z\, \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz + \int_{0}^{\infty} z\, \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz. \end{aligned} \tag{3.2.1}$$

But using the substitution $w = -z$, we see that

$$\int_{-\infty}^{0} z\, \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \int_{0}^{\infty} (-w) \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw.$$

Then the two integrals in (3.2.1) cancel each other out, and leaving us with $E(Z) = 0$. ∎

As with discrete variables, means of absolutely continuous random variables can also be infinite or undefined.

EXAMPLE 3.2.5
Let X have density function given by

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

Then

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_{1}^{\infty} x\, (1/x^2)\, dx = \int_{1}^{\infty} (1/x)\, dx = \log x \Big|_{x=1}^{x=\infty} = \infty.$$

Hence, the expected value of X is infinite. ∎

EXAMPLE 3.2.6
Let Y have density function given by

$$f_Y(y) = \begin{cases} 1/2y^2 & y \geq 1 \\ 1/2y^2 & y \leq -1 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \int_{-\infty}^{-1} y(1/y^2)\, dy + \int_{1}^{\infty} y(1/y^2)\, dy \\ &= -\int_{1}^{\infty} (1/y)\, dy + \int_{1}^{\infty} (1/y)\, dy = -\infty + \infty, \end{aligned}$$

which is undefined. Hence, the expected value of Y is undefined (i.e., does not exist) in this case. ∎

Theorem 3.1.1 remains true in the continuous case, as follows.

Theorem 3.2.1
(a) Let X be an absolutely continuous random variable, with density function f_X, and let $g : R^1 \to R^1$ be some function. Then when the expectation of $g(X)$ exists,

$$E\left(g(X)\right) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx.$$

(b) Let X and Y be jointly absolutely continuous random variables, with joint density function $f_{X,Y}$, and let $h : R^2 \to R^1$ be some function. Then when the expectation of $h(X, Y)$ exists,

$$E\left(h(X, Y)\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

We do not prove Theorem 3.2.1 here; however, we shall use it often. For a first use of this result, we prove that expected values for absolutely continuous random variables are still linear.

Theorem 3.2.2 (*Linearity of expected values*) Let X and Y be jointly absolutely continuous random variables, and let a and b be real numbers. Then $E(aX+bY) = aE(X) + bE(Y)$.

PROOF Let $f_{X,Y}$ be the joint density function of X and Y. Then using Theorem 3.2.1, we compute that

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by)\, f_{X,Y}\,(x, y)\, dx\, dy \\ &= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}\,(x, y)\, dx\, dy + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}\,(x, y)\, dx\, dy \\ &= a \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{X,Y}\,(x, y)\, dy \right) dx \\ &\quad + b \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{X,Y}\,(x, y)\, dx \right) dy. \end{aligned}$$

But $\int_{-\infty}^{\infty} f_{X,Y}\,(x, y)\, dy = f_X(x)$ and $\int_{-\infty}^{\infty} f_{X,Y}\,(x, y)\, dx = f_Y(y)$, so

$$E(Z) = a \int_{-\infty}^{\infty} x f_X(x)\, dx + b \int_{-\infty}^{\infty} y f_Y(y)\, dy = aE(X) + bE(Y),$$

as claimed. ∎

Just as in the discrete case, we have that $E(X + c) = E(X) + c$ for an absolutely continuous random variable X. Note, however, that this is not implied by Theorem 3.2.2 because the constant c is a discrete, not absolutely continuous, random variable.

In fact, we need a more general treatment of expectation to obtain this result (see Section 3.7). In any case, the result is true and we again have that the mean of a random variable serves as a measure of the location of the probability distribution of X. In Figure 3.2.1, we have plotted the densities and means of the absolutely continuous random variables X and $X + c$. The change of variable results from Section 2.6.2 give $f_{X+c}(x) = f_X(x - c)$.

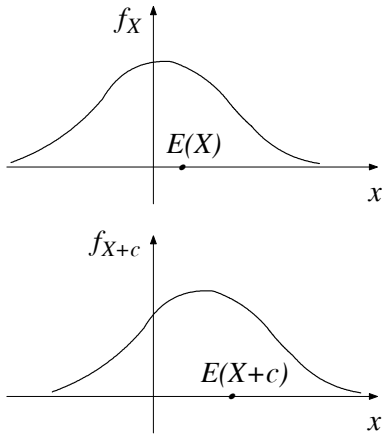

Figure 3.2.1: The densities and means of absolutely continuous random variables X and $X + c$.

EXAMPLE 3.2.7 *The $N(\mu, \sigma^2)$ Distribution*
Let $X \sim N(\mu, \sigma^2)$. Then we know (cf. Exercise 2.6.3) that if $Z = (X - \mu)/\sigma$, then $Z \sim N(0, 1)$. Hence, we can write $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$. But we know (see Example 3.2.4) that $E(Z) = 0$ and (Example 3.1.5) that $E(\mu) = \mu$. Hence, using Theorem 3.2.2, $E(X) = E(\mu + \sigma Z) = E(\mu) + \sigma E(Z) = \mu + \sigma(0) = \mu$. ∎

If X and Y are *independent*, then the following results show that we again have $E(XY) = E(X)E(Y)$.

Theorem 3.2.3 Let X and Y be jointly absolutely continuous random variables that are independent. Then $E(XY) = E(X)E(Y)$.

PROOF Independence implies (Theorem 2.8.3) that $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. Using this, along with Theorem 3.2.1, we compute

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x, y)\, dx\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x) f_Y(y)\, dx\, dy \\ &= \left(\int_{-\infty}^{\infty} x\, f_X(x)\, dx\right)\left(\int_{-\infty}^{\infty} y\, f_Y(y)\, dy\right) = E(X)E(Y), \end{aligned}$$

as claimed. ∎

The monotonicity property (Theorem 3.1.4) still holds as well.

Theorem 3.2.4 (*Monotonicity*) Let X and Y be jointly continuous random variables, and suppose that $X \leq Y$. Then $E(X) \leq E(Y)$.

PROOF Let $f_{X,Y}$ be the joint density function of X and Y. Because $X \leq Y$, the density $f_{X,Y}$ can be chosen so that $f_{X,Y}(x, y) = 0$ whenever $x > y$. Now let $Z = Y - X$. Then by Theorem 3.2.1(b),

$$E(Z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - x)\, f_{X,Y}(x, y)\, dx\, dy\,.$$

Because $f_{X,Y}(x, y) = 0$ whenever $x > y$, this implies that $E(Z) \geq 0$. But by Theorem 3.2.2, $E(Z) = E(Y) - E(X)$. Hence, $E(Y) - E(X) \geq 0$, so that $E(Y) \geq E(X)$. ∎

Summary of Section 3.2

- If X is absolutely continuous, then $E(X) = \int x\, f_X(x)\, dx$.
- The expected values of the uniform, exponential, and normal distributions were computed.
- Expectation for absolutely continuous random variables is linear and monotone.
- If X and Y are independent, then we still have $E(XY) = E(X)\, E(Y)$.

EXERCISES

3.2.1 Compute C and $E(X)$ when the density function of X is given by each of the following.

(a)
$$f_X(x) = \begin{cases} C & 5 \leq x \leq 9 \\ 0 & \text{otherwise} \end{cases}$$

(b)
$$f_X(x) = \begin{cases} C(x+1) & 6 \leq x \leq 8 \\ 0 & \text{otherwise} \end{cases}$$

(c)
$$f_X(x) = \begin{cases} Cx^4 & -5 \leq x \leq -2 \\ 0 & \text{otherwise} \end{cases}$$

3.2.2 Let X and Y have joint density

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

as in Examples 2.7.6 and 2.7.7. Compute each of the following.

(a) $E(X)$

(b) $E(Y)$
(c) $E(3X + 7Y)$
(d) $E(X^2)$
(e) $E(Y^2)$
(f) $E(XY)$
(g) $E(XY + 14)$

3.2.3 Let X and Y have joint density

$$f_{X,Y}(x, y) = \begin{cases} (4xy + 3x^2y^2)/18 & 0 \le x \le 1, 0 \le y \le 3 \\ 0 & \text{otherwise.} \end{cases}$$

Compute each of the following.
(a) $E(X)$
(b) $E(Y)$
(c) $E(X^2)$
(d) $E(Y^2)$
(e) $E(Y^4)$
(f) $E(X^2Y^3)$

3.2.4 Let X and Y have joint density

$$f_{X,Y}(x, y) = \begin{cases} 6xy + (9/2)x^2y^2 & 0 \le y \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Compute each of the following.
(a) $E(X)$
(b) $E(Y)$
(c) $E(X^2)$
(d) $E(Y^2)$
(e) $E(Y^4)$
(f) $E(X^2Y^3)$

3.2.5 Let $X \sim$ Uniform$[3, 7]$ and $Y \sim$ Exponential(9). Compute $E(-5X - 6Y)$.

3.2.6 Let $X \sim$ Uniform$[-12, -9]$ and $Y \sim N(-8,\ 9)$. Compute $E(11X + 14Y + 3)$.

3.2.7 Let $Y \sim$ Exponential(9) and $Z \sim$ Exponential(8). Compute $E(Y + Z)$.

3.2.8 Let $Y \sim$ Exponential(9) and $Z \sim$ Gamma$(5,\ 4)$. Compute $E(Y + Z)$. (You may use Problem 3.2.16 below.)

3.2.9 Suppose X has density function $f(x) = 3/20(x^2 + x^3)$ for $0 < x < 2$, otherwise $f(x) = 0$. Compute each of $E(X)$, $E(X^2)$, and $E(X^3)$, and rank them from largest to smallest.

3.2.10 Suppose X has density function $f(x) = 12/7(x^2 + x^3)$ for $0 < x < 1$, otherwise $f(x) = 0$. Compute each of $E(X)$, $E(X^2)$, and $E(X^3)$ and rank them from largest to smallest.

3.2.11 Suppose men's heights (in centimeters) follow the distribution $N(174,\ 20^2)$, while those of women follow the distribution $N(160,\ 15^2)$. Compute the mean total height of a man–woman married couple.

3.2.12 Suppose X and Y are independent, with $E(X) = 5$ and $E(Y) = 6$. For each of the following variables Z, either compute $E(Z)$ or explain why we cannot determine

$E(Z)$ from the available information:
(a) $Z = X + Y$
(b) $Z = XY$
(c) $Z = 2X - 4Y$
(d) $Z = 2X(3 + 4Y)$
(e) $Z = (2 + X)(3 + 4Y)$
(f) $Z = (2 + X)(3X + 4Y)$

3.2.13 Suppose darts are randomly thrown at a wall. Let X be the distance (in centimeters) from the left edge of the dart's point to the left end of the wall, and let Y be the distance from the right edge of the dart's point to the left end of the wall. Assume the dart's point is 0.1 centimeters thick, and that $E(X) = 214$. Compute $E(Y)$.

3.2.14 Let X be the mean height of all citizens measured from the top of their head, and let Y be the mean height of all citizens measured from the top of their head or hat (whichever is higher). Must we have $E(Y) \geq E(X)$? Why or why not?

3.2.15 Suppose basketball teams A and B each have five players and that each member of team A is being "guarded" by a unique member of team B. Suppose it is noticed that each member of team A is taller than the corresponding gusrf from team B. Does it necessarily follow that the mean height of team A is larger than the mean height of team B? Why or why not?

PROBLEMS

3.2.16 Let $\alpha > 0$ and $\lambda > 0$, and let $X \sim \text{Gamma}(\alpha, \lambda)$. Prove that $E(X) = \alpha/\lambda$. (Hint: The computations are somewhat similar to those of Problem 2.4.15. You will also need property (2.4.7) of the gamma function.)

3.2.17 Suppose that X follows the logistic distribution (see Problem 2.4.18). Prove that $E(X) = 0$.

3.2.18 Suppose that X follows the Weibull(α) distribution (see Problem 2.4.19). Prove that $E(X) = \Gamma\left(\alpha^{-1} + 1\right)$.

3.2.19 Suppose that X follows the Pareto(α) distribution (see Problem 2.4.20) for $\alpha > 1$. Prove that $E(X) = 1/(\alpha - 1)$. What is $E(X)$ when $0 < \alpha \leq 1$?

3.2.20 Suppose that X follows the Cauchy distribution (see Problem 2.4.21). Argue that $E(X)$ does not exist. (Hint: Compute the integral in two parts, where the integrand is positive and where the integrand is negative.)

3.2.21 Suppose that X follows the Laplace distribution (see Problem 2.4.22). Prove that $E(X) = 0$.

3.2.22 Suppose that X follows the Beta(a, b) distribution (see Problem 2.4.24). Prove that $E(X) = a/(a + b)$.

3.2.23 Suppose that $(X_1, X_2) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ (see Problem 2.7.17). Prove that $E(X_i) = \alpha_i/(\alpha_1 + \alpha_2 + \alpha_3)$.

3.3 Variance, Covariance, and Correlation

Now that we understand expected value, we can use it to define various other quantities of interest. The numerical values of these quantities provide information about the distribution of random variables.

Given a random variable X, we know that the average value of X will be $E(X)$. However, this tells us nothing about how far X tends to be from $E(X)$. For that, we have the following definition.

Definition 3.3.1 The *variance* of a random variable X is the quantity

$$\sigma_x^2 = \text{Var}(X) = E\left((X - \mu_X)^2\right), \tag{3.3.1}$$

where $\mu_X = E(X)$ is the mean of X.

We note that it is also possible to write (3.3.1) as $\text{Var}(X) = E\left((X - E(X))^2\right)$; however, the multiple uses of "E" may be confusing. Also, because $(X - \mu_X)^2$ is always nonnegative, its expectation is always defined, so the variance of X is always defined.

Intuitively, the variance $\text{Var}(X)$ is a measure of how *spread out* the distribution of X is, or how *random* X is, or how much X *varies*, as the following example illustrates.

EXAMPLE 3.3.1
Let X and Y be two discrete random variables, with probability functions

$$p_X(x) = \begin{cases} 1 & x = 10 \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_Y(y) = \begin{cases} 1/2 & y = 5 \\ 1/2 & y = 15 \\ 0 & \text{otherwise,} \end{cases}$$

respectively.

Then $E(X) = E(Y) = 10$. However,

$$\text{Var}(X) = (10 - 10)^2(1) = 0,$$

while

$$\text{Var}(Y) = (5 - 10)^2(1/2) + (15 - 10)^2(1/2) = 25.$$

We thus see that, while X and Y have the same expected value, the variance of Y is much greater than that of X. This corresponds to the fact that Y is more random than X; that is, it varies more than X does. ∎

EXAMPLE 3.3.2
Let X have probability function given by

$$p_X(x) = \begin{cases} 1/2 & x = 2 \\ 1/6 & x = 3 \\ 1/6 & x = 4 \\ 1/6 & x = 5 \\ 0 & \text{otherwise.} \end{cases}$$

Then $E(X) = (2)(1/2) + (3)(1/6) + (4)(1/6) + (5)(1/6) = 3$. Hence,

$$\text{Var}(X) = ((2-3)^2)\frac{1}{2} + ((3-3)^2)\frac{1}{6} + ((4-3)^2)\frac{1}{6} + ((5-3)^2)\frac{1}{6} = 4/3. \blacksquare$$

EXAMPLE 3.3.3
Let $Y \sim \text{Bernoulli}(\theta)$. Then $E(Y) = \theta$. Hence,

$$\begin{aligned} \text{Var}(Y) &= E((Y-\theta)^2) = ((1-\theta)^2)(\theta) + ((0-\theta)^2)(1-\theta) \\ &= \theta - 2\theta^2 + \theta^3 + \theta^2 - \theta^3 = \theta - \theta^2 = \theta(1-\theta). \blacksquare \end{aligned}$$

The square in (3.3.1) implies that the "scale" of $\text{Var}(X)$ is different from the scale of X. For example, if X were measuring a distance in meters (m), then $\text{Var}(X)$ would be measuring in meters squared (m^2). If we then switched from meters to feet, we would have to multiply X by about 3.28084 but would have to multiply $\text{Var}(X)$ by about $(3.28084)^2$.

To correct for this "scale" problem, we can simply take the square root, as follows.

Definition 3.3.2 The *standard deviation* of a random variable X is the quantity

$$\sigma_X = \text{Sd}(X) = \sqrt{\text{Var}(X)} = \sqrt{E((X-\mu_X)^2)}.$$

It is reasonable to ask why, in (3.3.1), we need the square at all. Now, if we simply *omitted* the square and considered $E((X-\mu_X))$, we would always get zero (because $\mu_X = E(X)$), which is useless. On the other hand, we could instead use $E(|X-\mu_X|)$. This would, like (3.3.1), be a valid measure of the average distance of X from μ_X. Furthermore, it would not have the "scale problem" that $\text{Var}(X)$ does. However, we shall see that $\text{Var}(X)$ has many convenient properties. By contrast, $E(|X-\mu_X|)$ is very difficult to work with. Thus, it is purely for *convenience* that we define variance by $E((X-\mu_X)^2)$ instead of $E(|X-\mu_X|)$.

Variance will be very important throughout the remainder of this book. Thus, we pause to present some important properties of Var.

Theorem 3.3.1 Let X be any random variable, with expected value $\mu_X = E(X)$, and variance $\text{Var}(X)$. Then the following hold true:
(a) $\text{Var}(X) \geq 0$.
(b) If a and b are real numbers, $\text{Var}(aX+b) = a^2\,\text{Var}(X)$.
(c) $\text{Var}(X) = E(X^2) - (\mu_X)^2 = E(X^2) - E(X)^2$. (That is, variance is equal to the second moment minus the square of the first moment.)
(d) $\text{Var}(X) \leq E(X^2)$.

PROOF (a) This is immediate, because we always have $(X - \mu_X)^2 \geq 0$.
(b) We note that $\mu_{aX+b} = E(aX + b) = aE(X) + b = a\mu_X + b$, by linearity. Hence, again using linearity,

$$\begin{aligned} \text{Var}(aX + b) &= E\left((aX + b - \mu_{aX+b})^2\right) = E\left((aX + b - a\mu_X + b)^2\right) \\ &= a^2 E\left((X - \mu_X)^2\right) = a^2 \,\text{Var}(X). \end{aligned}$$

(c) Again, using linearity,

$$\begin{aligned} \text{Var}(X) &= E\left((X - \mu_X)^2\right) = E\left(X^2 - 2X\mu_X + (\mu_X)^2\right) \\ &= E(X^2) - 2E(X)\mu_X + (\mu_X)^2 = E(X^2) - 2(\mu_X)^2 + (\mu_X)^2 \\ &= E(X^2) - (\mu_X)^2. \end{aligned}$$

(d) This follows immediately from part (c) because we have $-(\mu_X)^2 \leq 0$. ∎

Theorem 3.3.1 often provides easier ways of computing variance, as in the following examples.

EXAMPLE 3.3.4 *Variance of the Exponential(λ) Distribution*
Let $W \sim$ Exponential(λ), so that $f_W(w) = \lambda e^{-\lambda w}$. Then $E(W) = 1/\lambda$. Also, using integration by parts,

$$\begin{aligned} E(W^2) &= \int_0^\infty w^2 \lambda e^{-\lambda w} dw = \int_0^\infty 2we^{-\lambda w} dw \\ &= (2/\lambda) \int_0^\infty w\lambda e^{-\lambda w} dw = (2/\lambda)E(W) = 2/\lambda^2. \end{aligned}$$

Hence, by part (c) of Theorem 3.3.1,

$$\text{Var}(W) = E(W^2) - (E(W))^2 = (2/\lambda^2) - (1/\lambda)^2 = 1/\lambda^2. \text{ ∎}$$

EXAMPLE 3.3.5
Let $W \sim$ Exponential(λ), and let $Y = 5W + 3$. Then from the above example, $\text{Var}(W) = 1/\lambda^2$. Then, using part (b) of Theorem 3.3.1,

$$\text{Var}(Y) = \text{Var}(5W + 3) = 25 \,\text{Var}(W) = 25/\lambda^2. \text{ ∎}$$

Because $\sqrt{a^2} = |a|$, part (b) of Theorem 3.3.1 immediately implies a corresponding fact about standard deviation.

Corollary 3.3.1 Let X be any random variable, with standard deviation $\text{Sd}(X)$, and let a be any real number. Then $\text{Sd}(aX) = |a| \,\text{Sd}(X)$.

EXAMPLE 3.3.6
Let $W \sim$ Exponential(λ), and let $Y \sim 5W + 3$. Then using the above examples, we see that $\text{Sd}(W) = (\text{Var}(W))^{1/2} = \left(1/\lambda^2\right)^{1/2} = 1/\lambda$. Also, $\text{Sd}(Y) = (\text{Var}(Y))^{1/2} = \left(25/\lambda^2\right)^{1/2} = 5/\lambda$. This agrees with Corollary 3.3.1, since $\text{Sd}(Y) = |5| \,\text{Sd}(W)$. ∎

EXAMPLE 3.3.7 *Variance and Standard Deviation of the $N(\mu, \sigma^2)$ Distribution*
Suppose that $X \sim N(\mu, \sigma^2)$. In Example 3.2.7 we established that $E(X) = \mu$. Now we compute $\text{Var}(X)$.

First consider $Z \sim N(0, 1)$. Then from Theorem 3.3.1(c) we have that

$$\text{Var}(Z) = E\left(Z^2\right) = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz.$$

Then, putting $u = z$, $dv = z \exp\left\{-z^2/2\right\}$ (so $du = 1$, $v = -\exp\left\{-z^2/2\right\}$), and using integration by parts, we obtain

$$\text{Var}(Z) = -\frac{1}{\sqrt{2\pi}} z \exp\left\{-z^2/2\right\}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz = 1$$

and $\text{Sd}(Z) = 1$.

Now, for $\sigma > 0$, put $X = \mu + \sigma Z$. We then have $X \sim N(\mu, \sigma^2)$. From Theorem 3.3.1(b) we have that

$$\text{Var}(X) = \text{Var}(\mu + \sigma Z) = \sigma^2 \text{Var}(Z) = \sigma^2$$

and $\text{Sd}(X) = \sigma$. This establishes the variance of the $N(\mu, \sigma^2)$ distribution as σ^2 and the standard deviation as σ.

In Figure 3.3.1, we have plotted three normal distributions, all with mean 0 but different variances.

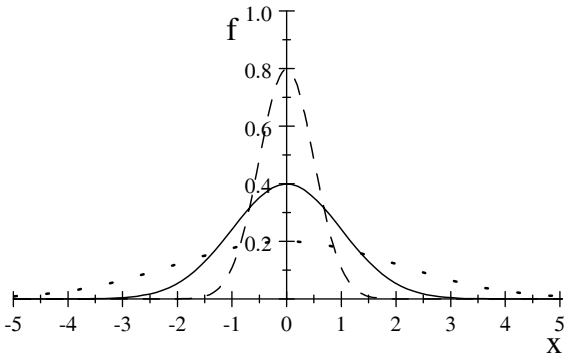

Figure 3.3.1: Plots of the the $N(0, 1)$ (solid line), the $N(0, 1/4)$ (dashed line) and the $N(0, 4)$ (dotted line) density functions.

The effect of the variance on the amount of spread of the distribution about the mean is quite clear from these plots. As σ^2 increases, the distribution becomes more diffuse; as it decreases, it becomes more concentrated about the mean 0. ∎

So far we have considered the variance of one random variable at a time. However, the related concept of covariance measures the *relationship* between two random variables.

Definition 3.3.3 The *covariance* of two random variables X and Y is given by

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

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

EXAMPLE 3.3.8
Let X and Y be discrete random variables, with joint probability function $p_{X,Y}$ given by

$$p_{X,Y}(x, y) = \begin{cases} 1/2 & x = 3, y = 4 \\ 1/3 & x = 3, y = 6 \\ 1/6 & x = 5, y = 6 \\ 0 & \text{otherwise.} \end{cases}$$

Then $E(X) = (3)(1/2) + (3)(1/3) + (5)(1/6) = 10/3$, and $E(Y) = (4)(1/2) + (6)(1/3) + (6)(1/6) = 5$. Hence,

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E((X - 10/3)(Y - 5)) \\ &= (3 - 10/3)(4 - 5)/2 + (3 - 10/3)(6 - 5)/3 + (5 - 10/3)(6 - 5)/6 \\ &= 1/3. \blacksquare \end{aligned}$$

EXAMPLE 3.3.9
Let X be any random variable with $\mathrm{Var}(X) > 0$. Let $Y = 3X$, and let $Z = -4X$. Then $\mu_Y = 3\mu_X$ and $\mu_Z = -4\mu_X$. Hence,

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E((X - \mu_X)(Y - \mu_Y)) = E((X - \mu_X)(3X - 3\mu_X)) \\ &= 3\,E((X - \mu_X)^2) = 3\,\mathrm{Var}(X), \end{aligned}$$

while

$$\begin{aligned} \mathrm{Cov}(X, Z) &= E((X - \mu_X)(Z - \mu_Z)) = E((X - \mu_X)((-4)X - (-4)\mu_X)) \\ &= (-4)E((X - \mu_X)^2) = -4\,\mathrm{Var}(X). \end{aligned}$$

Note in particular that $\mathrm{Cov}(X, Y) > 0$, while $\mathrm{Cov}(X, Z) < 0$. Intuitively, this says that Y increases when X increases, whereas Z decreases when X increases. $\blacksquare$

We begin with some simple facts about covariance. Obviously, we always have $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$. We also have the following result.

Theorem 3.3.2 (*Linearity of covariance*) Let X, Y, and Z be three random variables. Let a and b be real numbers. Then

$$\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z).$$

PROOF Note that by linearity, $\mu_{aX+bY} \equiv E(aX + bY) = aE(X) + bE(Y) \equiv a\mu_X + b\mu_Y$. Hence,

$$\begin{aligned}
\text{Cov}(aX + bY, Z) &= E\left((aX + bY - \mu_{aX+bY})(Z - \mu_Z)\right) \\
&= E\left((aX + bY - a\mu_X - b\mu_Y)(Z - \mu_Z)\right) \\
&= E\left((aX - a\mu_X + bY - b\mu_Y)(Z - \mu_Z)\right) \\
&= aE\left((X - \mu_X)(Z - \mu_Z)\right) + bE\left((Y - \mu_Y)(Z - \mu_Z)\right) \\
&= a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z),
\end{aligned}$$

and the result is established. ∎

We also have the following identity, which is similar to Theorem 3.3.1(c).

Theorem 3.3.3 Let X and Y be two random variables. Then

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

PROOF Using linearity, we have

$$\begin{aligned}
\text{Cov}(X, Y) &= E\left((X - \mu_X)(Y - \mu_Y)\right) = E\left(XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y\right) \\
&= E(XY) - \mu_X E(Y) - E(X)\mu_Y + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y. \blacksquare
\end{aligned}$$

Corollary 3.3.2 If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

PROOF Because X and Y are independent, we know (Theorems 3.1.3 and 3.2.3) that $E(XY) = E(X)\,E(Y)$. Hence, the result follows immediately from Theorem 3.3.3. ∎

We note that the converse to Corollary 3.3.2 is false, as the following example shows.

EXAMPLE 3.3.10 *Covariance* 0 *Does Not Imply Independence.*
Let X and Y be discrete random variables, with joint probability function $p_{X,Y}$ given by

$$p_{X,Y}(x, y) = \begin{cases} 1/4 & x = 3,\ y = 5 \\ 1/4 & x = 4,\ y = 9 \\ 1/4 & x = 7,\ y = 5 \\ 1/4 & x = 6,\ y = 9 \\ 0 & \text{otherwise.} \end{cases}$$

Then $E(X) = (3)(1/4) + (4)(1/4) + (7)(1/4) + (6)(1/4) = 5$, $E(Y) = (5)(1/4) + (9)(1/4) + (5)(1/4) + (9)(1/4) = 7$, and $E(XY) = (3)(5)(1/4) + (4)(9)(1/4) + (7)(5)(1/4) + (6)(9)(1/4) = 35$. We obtain $\text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = 35 - (5)(7) = 0$.

On the other hand, X and Y are clearly not independent. For example, $P(X = 4) > 0$ and $P(Y = 5) > 0$, but $P(X = 4, Y = 5) = 0$, so $P(X = 4, Y = 5) \neq P(X = 4)\,P(Y = 5)$. ∎

There is also an important relationship between variance and covariance.

Theorem 3.3.4
(a) For any random variables X and Y,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y).$$

(b) More generally, for any random variables $X_1, \ldots, X_n$,

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

PROOF We prove part (b) here; part (a) then follows as the special case $n = 2$.
Note that by linearity,

$$\mu_{\sum_i X_i} \equiv E\left(\sum_i X_i\right) = \sum_i E(X_i) \equiv \sum_i \mu_{X_i}.$$

Therefore, we have that

$$\begin{aligned}
&\text{Var}\left(\sum_i X_i\right) \\
&= E\left(\left(\sum_i X_i - \mu_{\sum_i X_i}\right)^2\right) = E\left(\left(\sum_i X_i - \sum_i \mu_i\right)^2\right) \\
&= E\left(\left(\sum_i (X_i - \mu_i)\right)^2\right) = E\left(\left(\sum_i (X_i - \mu_i)\right)\left(\sum_j (X_j - \mu_j)\right)\right) \\
&= E\left(\sum_{i,j} (X_i - \mu_i)\,(X_j - \mu_j)\right) = \sum_{i,j} E((X_i - \mu_i)\,(X_j - \mu_j)) \\
&= \sum_{i=j} E((X_i - \mu_i)\,(X_j - \mu_j)) + 2\sum_{i<j} E((X_i - \mu_i)\,(X_j - \mu_j)) \\
&= \sum_i \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j). \blacksquare
\end{aligned}$$

Combining Theorem 3.3.4 with Corollary 3.3.2, we obtain the following.

Corollary 3.3.3
(a) If X and Y are independent, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
(b) If $X_1, \ldots, X_n$ are independent, then $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$.

One use of Corollary 3.3.3 is the following.

EXAMPLE 3.3.11
Let $Y \sim \text{Binomial}(n, \theta)$. What is $\text{Var}(Y)$? Recall that we can write

$$Y = X_1 + X_2 + \cdots + X_n,$$

where the X_i are independent, with $X_i \sim \text{Bernoulli}(\theta)$. We have already seen that $\text{Var}(X_i) = \theta(1-\theta)$. Hence, from Corollary 3.3.3,

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) \\ &= \theta(1-\theta) + \theta(1-\theta) + \cdots + \theta(1-\theta) = n\theta(1-\theta). \blacksquare \end{aligned}$$

Another concept very closely related to covariance is correlation.

Definition 3.3.4 The *correlation* of two random variables X and Y is given by

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{Sd}(X)\,\text{Sd}(Y)} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

provided $\text{Var}(X) < \infty$ and $\text{Var}(Y) < \infty$.

EXAMPLE 3.3.12
As in Example 3.3.2, let X be any random variable with $\text{Var}(X) > 0$, let $Y = 3X$, and let $Z = -4X$. Then $\text{Cov}(X, Y) = 3\,\text{Var}(X)$ and $\text{Cov}(X, Z) = -4\,\text{Var}(X)$. But by Corollary 3.3.1, $\text{Sd}(Y) = 3\,\text{Sd}(X)$ and $\text{Sd}(Z) = 4\,\text{Sd}(X)$. Hence,

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{Sd}(X)\,\text{Sd}(Y)} = \frac{3\,\text{Var}(X)}{\text{Sd}(X)3\,\text{Sd}(X)} = \frac{\text{Var}(X)}{\text{Sd}(X)^2} = 1,$$

because $\text{Sd}(X)^2 = \text{Var}(X)$. Also, we have that

$$\text{Corr}(X, Z) = \frac{\text{Cov}(X, Z)}{\text{Sd}(X)\,\text{Sd}(Z)} = \frac{-4\,\text{Var}(X)}{\text{Sd}(X)4\,\text{Sd}(X)} = -\frac{\text{Var}(X)}{\text{Sd}(X)^2} = -1.$$

Intuitively, this again says that Y increases when X increases, whereas Z decreases when X increases. However, note that the scale factors 3 and -4 have cancelled out; only their signs were important. ∎

We shall see later, in Section 3.6, that we always have $-1 \leq \text{Corr}(X, Y) \leq 1$, for any random variables X and Y. Hence, in Example 3.3.12, Y has the largest possible correlation with X (which makes sense because Y increases whenever X does, without exception), while Z has the smallest possible correlation with X (which makes sense because Z decreases whenever X does). We will also see that $\text{Corr}(X, Y)$ is a measure of the extent to which a linear relationship exists between X and Y.

EXAMPLE 3.3.13 *The Bivariate Normal*$(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ *Distribution*
We defined this distribution in Example 2.7.9. It turns out that when (X, Y) follows this joint distribution then, (from Problem 2.7.13) $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. Further, we have that (see Problem 3.3.17) $\text{Corr}(X, Y) = \rho$. In the following graphs,

we have plotted samples of $n = 1000$ values of (X, Y) from bivariate normal distributions with $\mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$, and various values of ρ. Note that we used (2.7.1) to generate these samples.

From these plots we can see the effect of ρ on the joint distribution. Figure 3.3.2 shows that when $\rho = 0$, the point cloud is roughly circular. It becomes elliptical in Figure 3.3.3 with $\rho = 0.5$, and more tightly concentrated about a line in Figure 3.3.4 with $\rho = 0.9$. As we will see in Section 3.6, the points will lie exactly on a line when $\rho = 1$.

Figure 3.3.5 demonstrates the effect of a negative correlation. With positive correlations, the value of Y tends to increase with X, as reflected in the upward slope of the point cloud. With negative correlations, Y tends to decrease with X, as reflected in the negative slope of the point cloud. ∎

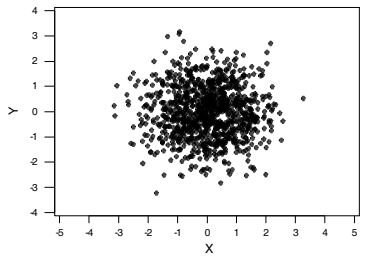

Figure 3.3.2: A sample of $n = 1000$ values (X, Y) from the Bivariate Normal $(0, 0, 1, 1, 0)$ distribution.

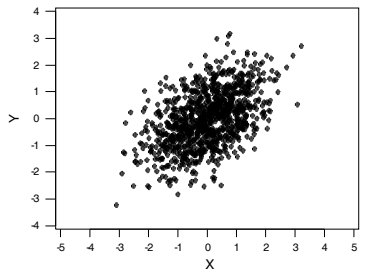

Figure 3.3.3: A sample of $n = 1000$ values (X, Y) from the Bivariate Normal $(0, 0, 1, 1, 0.5)$ distribution.

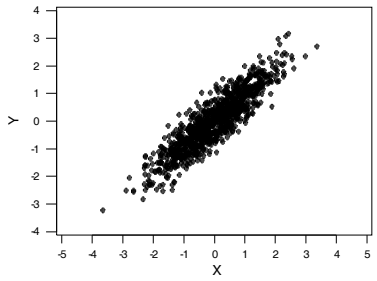

Figure 3.3.4: A sample of $n = 1000$ values (X, Y) from the Bivariate Normal $(0, 0, 1, 1, 0.9)$ distribution.

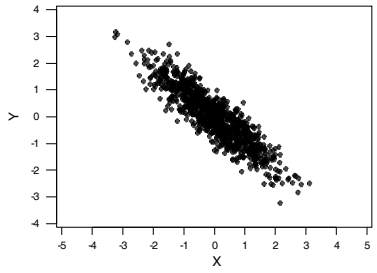

Figure 3.3.5: A sample of $n = 1000$ values (X, Y) from the Bivariate Normal $(0, 0, 1, 1, -0.9)$ distribution.

Summary of Section 3.3

- The variance of a random variable X measures how far it tends to be from its mean and is given by $\text{Var}(X) = E((X - \mu_X)^2) = E(X^2) - (E(X))^2$.
- The variances of many standard distributions were computed.
- The standard deviation of X equals $\text{Sd}(X) = \sqrt{\text{Var}(X)}$.
- $\text{Var}(X) \geq 0$, and $\text{Var}(aX + b) = a^2 \text{Var}(X)$; also $\text{Sd}(aX + b) = |a| \text{Sd}(X)$.
- The covariance of random variables X and Y measures how they are related and is given by $\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_y)) = E(XY) - E(X)\,E(Y)$.
- If X and Y are independent, then $\text{Cov}(X, Y) = 0$.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$. If X and Y are independent, this equals $\text{Var}(X) + \text{Var}(Y)$.
- The correlation of X and Y is $\text{Corr}(X, Y) = \text{Cov}(X, Y)/(\text{Sd}(X)\,\text{Sd}(Y))$.

EXERCISES

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

$$p_{X,Y}(x, y) = \begin{cases} 1/2 & x = 3,\ y = 5 \\ 1/6 & x = 3,\ y = 9 \\ 1/6 & x = 6,\ y = 5 \\ 1/6 & x = 6,\ y = 9 \\ 0 & \text{otherwise,} \end{cases}$$

with $E(X) = 4$, $E(Y) = 19/3$, and $E(XY) = 26$, as in Example 3.1.16.
(a) Compute $\text{Cov}(X, Y)$.
(b) Compute $\text{Var}(X)$ and $\text{Var}(Y)$.
(c) Compute $\text{Corr}(X, Y)$.

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

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

as in Example 2.7.5.
(a) Compute $E(X)$ and $E(Y)$.
(b) Compute $\text{Cov}(X, Y)$.
(c) Compute $\text{Var}(X)$ and $\text{Var}(Y)$.
(d) Compute $\text{Corr}(X, Y)$.

3.3.3 Let X and Y have joint density

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

as in Exercise 3.2.2. Compute $\text{Corr}(X, Y)$.

3.3.4 Let X and Y have joint density

$$f_{X,Y}(x, y) = \begin{cases} 15x^3y^4 + 6x^2y^7 & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Compute $E(X)$, $E(Y)$, $\text{Var}(X)$, $\text{Var}(Y)$, $\text{Cov}(X, Y)$, and $\text{Corr}(X, Y)$.

3.3.5 Let Y and Z be two independent random variables, each with positive variance. Prove that $\text{Corr}(Y, Z) = 0$.

3.3.6 Let X, Y, and Z be three random variables, and suppose that X and Z are independent. Prove that $\text{Cov}(X + Y, Z) = \text{Cov}(Y, Z)$.

3.3.7 Let $X \sim \text{Exponential}(3)$ and $Y \sim \text{Poisson}(5)$. Assume X and Y are independent. Let $Z = X + Y$.
(a) Compute $\text{Cov}(X, Z)$.
(b) Compute $\text{Corr}(X, Z)$.

3.3.8 Prove that the variance of the Uniform$[L, R]$ distribution is given by the expression $(R-L)^2/12$.

3.3.9 Prove that $\text{Var}(X) = E(X(X-1)) - E(X)E(X-1)$. Use this to compute directly from the probability function that when $X \sim \text{Binomial}(n, \theta)$, then $\text{Var}(X) = n\theta(1-\theta)$.

3.3.10 Suppose you flip three fair coins. Let X be the number of heads showing, and let $Y = X^2$. Compute $E(X)$, $E(Y)$, $\text{Var}(X)$, $\text{Var}(Y)$, $\text{Cov}(X, Y)$, and $\text{Corr}(X, Y)$.

3.3.11 Suppose you roll two fair six-sided dice. Let X be the number showing on the first die, and let Y be the sum of the numbers showing on the two dice. Compute $E(X)$, $E(Y)$, $E(XY)$, and $\text{Cov}(X, Y)$.

3.3.12 Suppose you flip four fair coins. Let X be the number of heads showing, and let Y be the number of tails showing. Compute $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$.

3.3.13 Let X and Y be independent, with $X \sim \text{Bernoulli}(1/2)$ and $Y \sim \text{Bernoulli}(1/3)$. Let $Z = X + Y$ and $W = X - Y$. Compute $\text{Cov}(Z, W)$ and $\text{Corr}(Z, W)$.

3.3.14 Let X and Y be independent, with $X \sim \text{Bernoulli}(1/2)$ and $Y \sim N(0, 1)$. Let $Z = X + Y$ and $W = X - Y$. Compute $\text{Var}(Z)$, $\text{Var}(W)$, $\text{Cov}(Z, W)$, and $\text{Corr}(Z, W)$.

3.3.15 Suppose you roll one fair six-sided die and then flip as many coins as the number showing on the die. (For example, if the die shows 4, then you flip four coins.) Let X be the number showing on the die, and Y be the number of heads obtained. Compute $\text{Cov}(X, Y)$.

PROBLEMS

3.3.16 Let $X \sim N(0, 1)$, and let $Y = cX$.
(a) Compute $\lim_{c \searrow 0} \text{Cov}(X, Y)$.
(b) Compute $\lim_{c \nearrow 0} \text{Cov}(X, Y)$.
(c) Compute $\lim_{c \searrow 0} \text{Corr}(X, Y)$.
(d) Compute $\lim_{c \nearrow 0} \text{Corr}(X, Y)$.
(e) Explain why the answers in parts (c) and (d) are not the same.

3.3.17 Let X and Y have the bivariate normal distribution, as in Example 2.7.9. Prove that $\text{Corr}(X, Y) = \rho$. (Hint: Use (2.7.1).)

3.3.18 Prove that the variance of the Geometric(θ) distribution is given by $(1-\theta)/\theta^2$. (Hint: Use Exercise 3.3.9 and $\left((1-\theta)^x\right)'' = x(x-1)(1-\theta)^{x-2}$.)

3.3.19 Prove that the variance of the Negative-Binomial(r, θ) distribution is given by $r(1-\theta)/\theta^2$. (Hint: Use Problem 3.3.18.)

3.3.20 Let $\alpha > 0$ and $\lambda > 0$, and let $X \sim \text{Gamma}(\alpha, \lambda)$. Prove that $\text{Var}(X) = \alpha/\lambda^2$. (Hint: Recall Problem 3.2.16.)

3.3.21 Suppose that $X \sim \text{Weibull}(\alpha)$ distribution (see Problem 2.4.19). Prove that $\text{Var}(X) = \Gamma(2/\alpha + 1) - \Gamma^2(1/\alpha + 1)$. (Hint: Recall Problem 3.2.18.)

3.3.22 Suppose that $X \sim \text{Pareto}(\alpha)$ (see Problem 2.4.20) for $\alpha > 2$. Prove that $\text{Var}(X) = \alpha/((\alpha-1)^2(\alpha-2))$. (Hint: Recall Problem 3.2.19.)

3.3.23 Suppose that X follows the Laplace distribution (see Problem 2.4.22). Prove that $\text{Var}(X) = 2$. (Hint: Recall Problem 3.2.21.)

3.3.24 Suppose that $X \sim \text{Beta}(a, b)$ (see Problem 2.4.24). Prove that $\text{Var}(X) = ab/((a+b)^2(a+b+1))$. (Hint: Recall Problem 3.2.22.)

3.3.25 Suppose that $(X_1, X_2, X_3) \sim \text{Multinomial}(n, \theta_1, \theta_2, \theta_3)$. Prove that

$$\text{Var}(X_i) = n\theta_i(1-\theta_i),\ \text{Cov}(X_i, X_j) = -n\theta_i\theta_j,\ \text{when } i \neq j.$$

(Hint: Recall Problem 3.1.23.)

3.3.26 Suppose that $(X_1, X_2) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ (see Problem 2.7.17). Prove that

$$\begin{aligned} \text{Var}(X_i) &= \frac{\alpha_i\,(\alpha_1+\alpha_2+\alpha_3-\alpha_i)}{(\alpha_1+\alpha_2+\alpha_3)^2\,(\alpha_1+\alpha_2+\alpha_3+1)}, \\ \text{Cov}\,(X_1, X_2) &= \frac{-\alpha_1\alpha_2}{(\alpha_1+\alpha_2+\alpha_3)^2\,(\alpha_1+\alpha_2+\alpha_3+1)}. \end{aligned}$$

(Hint: Recall Problem 3.2.23.)

3.3.27 Suppose that $X \sim \text{Hypergeometric}(N, M, n)$. Prove that

$$\text{Var}(X) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}.$$

(Hint: Recall Problem 3.1.21 and use Exercise 3.3.9.)

3.3.28 Suppose you roll one fair six-sided die and then flip as many coins as the number showing on the die. (For example, if the die shows 4, then you flip four coins.) Let X be the number showing on the die, and Y be the number of heads obtained. Compute $\text{Corr}(X, Y)$.

CHALLENGES

3.3.29 Let Y be a nonnegative random variable. Prove that $E(Y) = 0$ if and only if $P(Y = 0) = 1$. (You may assume for simplicity that Y is discrete, but the result is true for any Y.)

3.3.30 Prove that $\text{Var}(X) = 0$ if and only if there is a real number c with $P(X = c) = 1$. (You may use the result of Challenge 3.3.29.)

3.3.31 Give an example of a random variable X, such that $E(X) = 5$, and $\text{Var}(X) = \infty$.

3.4 Generating Functions

Let X be a random variable. Recall that the cumulative distribution function of X, defined by $F_X(x) = P(X \leq x)$, contains all the information about the distribution of X (see Theorem 2.5.1). It turns out that there are other functions — the probability-generating function and the moment-generating function — that also provide information (sometimes all the information) about X and its expected values.

Definition 3.4.1 Let X be a random variable (usually discrete). Then we define its *probability-generating function*, r_X, by $r_X(t) = E(t^X)$ for $t \in R^1$.

Consider the following examples of probability-generating functions.

EXAMPLE 3.4.1 *The Binomial*(n, θ) *Distribution*
If $X \sim \text{Binomial}(n, \theta)$, then

$$\begin{aligned} r_X(t) &= E(t^X) = \sum_{i=0}^{n} P(X = i)t^i = \sum_{i=0}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} t^i \\ &= \sum_{i=0}^{n} \binom{n}{i} (t\theta)^i (1-\theta)^{n-i} = (t\theta + 1 - \theta)^n, \end{aligned}$$

using the binomial theorem. ∎

EXAMPLE 3.4.2 *The Poisson*(λ) *Distribution*
If $Y \sim \text{Poisson}(\lambda)$, then

$$\begin{aligned} r_Y(t) &= E(t^Y) = \sum_{i=0}^{\infty} P(Y = i)t^i = \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} t^i \\ &= \sum_{i=0}^{\infty} e^{-\lambda} \frac{(\lambda t)^i}{i!} = e^{-\lambda} e^{\lambda t} = e^{\lambda(t-1)}. \end{aligned}$$ ∎

The following theorem tells us that once we know the probability-generating function $r_X(t)$, then we can compute all the probabilities $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, etc.

Theorem 3.4.1 Let X be a discrete random variable, whose possible values are all nonnegative integers. Assume that $r_X(t_0) < \infty$ for some $t_0 > 0$. Then

$$\begin{aligned} r_X(0) &= P(X = 0), \\ r'_X(0) &= P(X = 1), \\ r''_X(0) &= 2\,P(X = 2), \end{aligned}$$

etc. In general,

$$r_X^{(k)}(0) = k!\,P(X = k),$$

where $r_X^{(k)}$ is the kth derivative of r_X.

PROOF Because the possible values are all nonnegative integers of the form $i = 0, 1, 2, \ldots$, we have

$$\begin{aligned} r_X(t) &= E(t^X) = \sum_x t^x P(X = x) = \sum_{i=0}^{\infty} t^i P(X = i) \\ &= t^0 P(X = 0) + t^1 P(X = 1) + t^2 P(X = 2) + t^3 P(X = 3) + \cdots, \end{aligned}$$

so that

$$r_X(t) = 1P(X = 0) + t^1 P(X = 1) + t^2 P(X = 2) + t^3 P(X = 3) + \cdots. \quad (3.4.1)$$

Substituting $t = 0$ into (3.4.1), every term vanishes except the first one, and we obtain $r_X(0) = P(X = 0)$. Taking derivatives of both sides of (3.4.1), we obtain

$$r'_X(t) = 1P(X = 1) + 2t^1 P(X = 2) + 3t^2 P(X = 3) + \cdots,$$

and setting $t = 0$ gives $r'_X(0) = P(X = 1)$. Taking another derivative of both sides gives

$$r''_X(t) = 2P(X = 2) + 3 \cdot 2\, t^1 P(X = 3) + \cdots$$

and setting $t = 0$ gives $r''_X(0) = 2\,P(X = 2)$. Continuing in this way, we obtain the general formula. ∎

We now apply Theorem 3.4.1 to the binomial and Poisson distributions.

EXAMPLE 3.4.3 *The Binomial*(n, θ) *Distribution*
From Example 3.4.1, we have that

$$\begin{aligned} r_X(0) &= (1-\theta)^n \\ r'_X(0) &= n(t\theta + 1 - \theta)^{n-1}(\theta)\,\Big|_{t=0} = n(1-\theta)^{n-1}\theta \\ r''_X(0) &= n(n-1)(t\theta + 1 - \theta)^{n-2}(\theta)(\theta)\,\Big|_{t=0} = n(n-1)(1-\theta)^{n-2}\theta^2, \end{aligned}$$

etc. It is thus verified directly that

$$\begin{aligned} P(X = 0) &= r_X(0) \\ P(X = 1) &= r'_X(0) \\ 2\,P(X = 2) &= r''_X(0), \end{aligned}$$

etc. ∎

EXAMPLE 3.4.4 *The Poisson*(λ) *Distribution*
From Example 3.4.2, we have that

$$\begin{aligned} r_X(0) &= e^{-\lambda} \\ r'_X(0) &= \lambda e^{-\lambda} \\ r''_X(0) &= \lambda^2 e^{-\lambda}, \end{aligned}$$

etc. It is again verified that

$$\begin{aligned} P(X = 0) &= r_X(0) \\ P(X = 1) &= r'_X(0) \\ 2\,P(X = 2) &= r''_X(0), \end{aligned}$$

etc. ∎

From Theorem 3.4.1, we can see why r_X is called the probability-generating function. For, at least in the discrete case with the distribution concentrated on the nonnegative integers, we can indeed generate the probabilities for X from r_X. We thus

see immediately that for a random variable X that takes values only in $\{0, 1, 2, \ldots\}$, r_X is unique. By this we mean that if X and Y are concentrated on $\{0, 1, 2, \ldots\}$ and $r_X = r_Y$, then X and Y have the same distribution. This uniqueness property of the probability-generating function can be very useful in trying to determine the distribution of a random variable that takes only values in $\{0, 1, 2, \ldots\}$.

It is clear that the probability-generating function tells us a lot — in fact, everything — about the distribution of random variables concentrated on the nonnegative integers. But what about other random variables? It turns out that there are other quantities, called moments, associated with random variables that are quite informative about their distributions.

Definition 3.4.2 Let X be a random variable, and let k be a positive integer. Then the kth *moment* of X is the quantity $E(X^k)$, provided this expectation exists.

Note that if $E(X^k)$ exists and is finite, it can be shown that $E(X^l)$ exists and is finite when $0 \leq l < k$.

The first moment is just the mean of the random variable. This can be taken as a measure of where the central mass of probability for X lies in the real line, at least when this distribution is unimodal (has a single peak) and is not too highly skewed. The second moment $E(X^2)$, together with the first moment, gives us the variance through $\text{Var}(X) = E(X^2) - (E(X))^2$. Therefore, the first two moments of the distribution tell us about the location of the distribution and the spread, or degree of concentration, of that distribution about the mean. In fact, the higher moments also provide information about the distribution.

Many of the most important distributions of probability and statistics have all of their moments finite; in fact, they have what is called a moment-generating function.

Definition 3.4.3 Let X be any random variable. Then its *moment-generating function* m_X is defined by $m_X(s) = E(e^{sX})$ at $s \in R^1$.

The following example computes the moment-generating function of a well-known distribution.

EXAMPLE 3.4.5 *The Exponential(λ) Distribution*
Let $X \sim \text{Exponential}(\lambda)$. Then for $s < \lambda$,

$$\begin{aligned} m_X(s) &= E(e^{sX}) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx = \int_0^{\infty} e^{sx} \lambda e^{-\lambda x}\, dx \\ &= \int_0^{\infty} \lambda\, e^{(s-\lambda)x}\, dx = \frac{\lambda\, e^{(s-\lambda)x}}{s-\lambda} \Big|_{x=0}^{x=\infty} = -\frac{\lambda e^{(s-\lambda)0}}{s-\lambda} \\ &= -\frac{\lambda}{s-\lambda} = \lambda(\lambda - s)^{-1}. \blacksquare \end{aligned}$$

A comparison of Definitions 3.4.1 and 3.4.3 immediately gives the following.

Theorem 3.4.2 Let X be any random variable. Then $m_X(s) = r_X(e^s)$.

This result can obviously help us evaluate some moment-generating functions when we have r_X already.

EXAMPLE 3.4.6
Let $Y \sim \text{Binomial}(n, \theta)$. Then we know that $r_Y(t) = (t\theta + 1 - \theta)^n$. Hence, $m_Y(s) = r_Y(e^s) = (e^s\theta + 1 - \theta)^n$. ∎

EXAMPLE 3.4.7
Let $Z \sim \text{Poisson}(\lambda)$. Then we know that $r_Z(t) = e^{\lambda(t-1)}$. Hence, $m_Z(s) = r_Z(e^s) = e^{\lambda(e^s-1)}$. ∎

The following theorem tells us that once we know the moment-generating function $m_X(t)$, we can compute all the moments $E(X)$, $E(X^2)$, $E(X^3)$, etc.

Theorem 3.4.3 Let X be any random variable. Suppose that for some $s_0 > 0$, it is true that $m_X(s) < \infty$ whenever $s \in (-s_0, s_0)$. Then

$$\begin{aligned} m_X(0) &= 1 \\ m'_X(0) &= E(X) \\ m''_X(0) &= E(X^2), \end{aligned}$$

etc. In general,

$$m_X^{(k)}(0) = E(X^k),$$

where $m_X^{(k)}$ is the kth derivative of m_X.

PROOF We know that $m_X(s) = E(e^{sX})$. We have

$$m_X(0) = E(e^{0X}) = E(e^0) = E(1) = 1.$$

Also, taking derivatives, we see[3] that $m'_X(s) = E(X\, e^{sX})$, so

$$m'_X(0) = E(X\, e^{0X}) = E(Xe^0) = E(X).$$

Taking derivatives again, we see that $m''_X(s) = E(X^2 e^{sX})$, so

$$m''_X(0) = E(X^2\, e^{0X}) = E(X^2 e^0) = E(X^2).$$

Continuing in this way, we obtain the general formula. ∎

We now consider an application of Theorem 3.4.3.

EXAMPLE 3.4.8 *The Mean and Variance of the Exponential(λ) Distribution*
Using the moment-generating function computed in Example 3.4.5, we have

$$m'_X(s) = (-1)\lambda(\lambda - s)^{-2}(-1) = \lambda(\lambda - s)^{-2}.$$

[3] Strictly speaking, interchanging the order of derivative and expectation is justified by analytic function theory and requires that $m_X(s) < \infty$ whenever $|s| < s_0$.

Therefore,

$$E(X) = m'_X(0) = \lambda(\lambda - 0)^{-2} = \lambda/\lambda^2 = 1/\lambda,$$

as it should. Also,

$$E(X^2) = m''_X(0) = (-2)\lambda(\lambda - 0)^{-3}(-1) = 2\lambda/\lambda^3 = 2/\lambda^2,$$

so we have

$$\text{Var}(X) = E(X^2) - (E(X)) = (2/\lambda^2) - (1/\lambda)^2 = 1/\lambda^2.$$

This provides an easy way of computing the variance of X. ∎

EXAMPLE 3.4.9 *The Mean and Variance of the Poisson(λ) Distribution*
In Example 3.4.7, we obtained $m_Z(s) = \exp(\lambda(e^s - 1))$. So we have

$$\begin{aligned} E(X) &= m'_X(0) = \lambda e^0 \exp\left(\lambda(e^0 - 1)\right) = \lambda \\ E(X^2) &= m''_X(0) = \lambda e^0 \exp\left(\lambda(e^0 - 1)\right) + \left(\lambda e^0\right)^2 \exp\left(\lambda(e^0 - 1)\right) = \lambda + \lambda^2. \end{aligned}$$

Therefore, $\text{Var}(X) = E(X^2) - (E(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$. ∎

Computing the moment-generating function of a normal distribution is also important, but it is somewhat more difficult.

Theorem 3.4.4 If $X \sim N(0, 1)$, then $m_X(s) = e^{s^2/2}$.

PROOF Because X has density $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$, we have that

$$\begin{aligned} m_X(s) &= E(e^{sX}) = \int_{-\infty}^{\infty} e^{sx}\phi(x)\,dx = \int_{-\infty}^{\infty} e^{sx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{sx-(x^2/2)}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-s)^2/2+(s^2/2)}\,dx \\ &= e^{s^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-s)^2/2}\,dx. \end{aligned}$$

Setting $y = x - s$ (so that $dy = dx$), this becomes (using Theorem 2.4.2)

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

as claimed. ∎

One useful property of both probability-generating and moment-generating functions is the following.

Theorem 3.4.5 Let X and Y be random variables that are independent. Then we have
(a) $r_{X+Y}(t) = r_X(t)\,r_Y(t)$, and
(b) $m_{X+Y}(t) = m_X(t)\,m_Y(t)$.

PROOF Because X and Y are independent, so are t^X and t^Y (by Theorem 2.8.5). Hence, we know (by Theorems 3.1.3 and 3.2.3) that $E(t^X t^Y) = E(t^X)\,E(t^Y)$. Using this, we have

$$r_{X+Y}(t) = E\left(t^{X+Y}\right) = E\left(t^X t^Y\right) = E(t^X)E(t^Y) = r_X(t)r_Y(t).$$

Similarly,

$$m_{X+Y}(t) = E\left(e^{t(X+Y)}\right) = E\left(e^{tX}e^{tY}\right) = E(e^{tX})\,E(e^{tY}) = m_X(t)m_Y(t). \blacksquare$$

EXAMPLE 3.4.10
Let $Y \sim \text{Binomial}(n, \theta)$. Then, as in Example 3.1.15, we can write

$$Y = X_1 + \cdots + X_n,$$

where the $\{X_i\}$ are i.i.d. with $X_i \sim \text{Bernoulli}(\theta)$. Hence, Theorem 3.4.5 says we must have $r_Y(t) = r_{X_1}(t)\, r_{X_2}(t) \cdots r_{X_n}(t)$. But for any i,

$$r_{X_i}(t) = \sum_x t^x P(X = x) = t^1\theta + t^0(1-\theta) = \theta t + 1 - \theta.$$

Hence, we must have

$$r_Y(t) = (\theta t + 1 - \theta)(\theta t + 1 - \theta) \cdots (\theta t + 1 - \theta) = (\theta t + 1 - \theta)^n,$$

as already verified in Example 3.4.1. ∎

Moment-generating functions, when defined in a neighborhood of 0, completely define a distribution in the following sense. (We omit the proof, which is advanced.)

Theorem 3.4.6 (*Uniqueness theorem*) Let X be a random variable, such that for some $s_0 > 0$, we have $m_X(s) < \infty$ whenever $s \in (-s_0, s_0)$. Then if Y is some other random variable with $m_Y(s) = m_X(s)$ whenever $s \in (-s_0, s_0)$, then X and Y have the same distribution.

Theorems 3.4.1 and 3.4.6 provide a powerful technique for identifying distributions. For example, if we determine that the moment-generating function of X is $m_X(t) = \exp\left(s^2/2\right)$, then we know, from Theorems 3.4.4 and 3.4.6, that $X \sim N(0, 1)$. We can use this approach to determine the distributions of some complicated random variables.

EXAMPLE 3.4.11
Suppose that $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ and that these random variables are independent. Consider the distribution of $Y = \sum_{i=1}^n X_i$.

When $n = 1$ we have (from Problem 3.4.15)

$$m_Y(s) = \exp\left\{\mu_1 s + \frac{\sigma_1^2 s^2}{2}\right\}.$$

Then, using Theorem 3.4.5, we have that

$$\begin{aligned} m_Y(s) &= \prod_{i=1}^{n} m_{X_i}(s) = \prod_{i=1}^{n} \exp\left\{\mu_i s + \frac{\sigma_i^2 s^2}{2}\right\} \\ &= \exp\left\{\left(\sum_{i=1}^{n} \mu_i\right) s + \frac{\left(\sum_{i=1}^{n} \sigma_i^2\right) s^2}{2}\right\}. \end{aligned}$$

From Problem 3.4.15, and applying Theorem 3.4.6, we have that

$$Y \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right). \blacksquare$$

Generating functions can also help us with compound distributions, which are defined as follows.

Definition 3.4.4 Let $X_1, X_2, \ldots$ be i.i.d., and let N be a nonnegative, integer-valued random variable which is independent of the $\{X_i\}$. Let

$$S = \sum_{i=1}^{N} X_i. \tag{3.4.2}$$

Then S is said to have a *compound distribution*.

A compound distribution is obtained from a sum of i.i.d. random variables, where the number of terms in the sum is randomly distributed independently of the terms in the sum. Note that $S = 0$ when $N = 0$. Such distributions have applications in areas like insurance, where the $X_1, X_2, \ldots$ are claims and N is the number of claims presented to an insurance company during a period. Therefore, S represents the total amount claimed against the insurance company during the period. Obviously, the insurance company wants to study the distribution of S, as this will help determine what it has to charge for insurance to ensure a profit.

The following theorem is important in the study of compound distributions.

Theorem 3.4.7 If S has a compound distribution as in (3.4.2), then
(a) $E(S) = E(X_1)E(N)$.
(b) $m_S(s) = r_N(m_{X_1}(s))$.

PROOF See Section 3.8 for the proof of this result. ∎

3.4.1 Characteristic Functions (Advanced)

One problem with moment-generating functions is that they can be *infinite* in any open interval about $s = 0$. Consider the following example.

EXAMPLE 3.4.12
Let X be a random variable having density

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

Then

$$m_X(s) = E(e^{sX}) = \int_1^\infty e^{sx}(1/x^2)\,dx.$$

For any $s > 0$, we know that e^{sx} grows faster than x^2, so that $\lim_{x\to\infty} e^{sx}/x^2 = \infty$. Hence, $m_X(s) = \infty$ whenever $s > 0$.

Does X have any finite moments? We have that

$$E(X) = \int_1^\infty x(1/x^2)\,dx = \int_1^\infty (1/x)\,dx = \ln x|_{x=1}^{x=\infty} = \infty,$$

so, in fact, the first moment does not exist. From this we conclude that X does not have any moments. ∎

The random variable X in the above example does *not* satisfy the condition of Theorem 3.4.3 that $m_X(s) < \infty$ whenever $|s| < s_0$, for some $s_0 > 0$. Hence, Theorem 3.4.3 (like most other theorems that make use of moment-generating functions) does not apply. There is, however, a similarly defined function that does not suffer from this defect, given by the following definition.

Definition 3.4.5 Let X be any random variable. Then we define its *characteristic function*, c_X, by

$$c_X(s) = E(e^{isX}) \tag{3.4.3}$$

for $s \in R^1$.

So the definition of c_X is just like the definition of m_X, except for the introduction of the imaginary number $i = \sqrt{-1}$. Using properties of complex numbers, we see that (3.4.3) can also be written as $c_X(s) = E(\cos(sX)) + i\,E(\sin(sX))$ for $s \in R^1$.

Consider the following examples.

EXAMPLE 3.4.13 *The Bernoulli Distribution*
Let $X \sim$ Bernoulli(θ). Then

$$\begin{aligned} c_X(s) &= E(e^{isX}) = (e^{is0})(1-\theta) + (e^{is1})(\theta) \\ &= (1)(1-\theta) + e^{is}(\theta) = 1 - \theta + \theta e^{is} \\ &= 1 - \theta + \theta\cos s + i\theta \sin s. \end{aligned}$$ ∎

EXAMPLE 3.4.14
Let X have probability function given by

$$p_X(x) = \begin{cases} 1/6 & x = 2 \\ 1/3 & x = 3 \\ 1/2 & x = 4 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} c_X(s) &= E(e^{isX}) = (e^{is2})(1/6) + (e^{is3})(1/3) + (e^{is4})(1/2) \\ &= (1/6)\cos 2s + (1/3)\cos 3s + (1/2)\cos 4s \\ &\quad +(1/6)i\sin 2s + (1/3)i\sin 3s + i(1/2)\sin 4s. \blacksquare \end{aligned}$$

EXAMPLE 3.4.15
Let Z have probability function given by

$$p_Z(z) = \begin{cases} 1/2 & z = 1 \\ 1/2 & z = -1 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} c_Z(s) &= E(e^{isZ}) = (e^{is})(1/2) + (e^{-is})(1/2) \\ &= (1/2)\cos(s) + (1/2)\cos(-s) + (1/2)\sin(s) + (1/2)\sin(-s) \\ &= (1/2)\cos s + (1/2)\cos s + (1/2)\sin s - (1/2)\sin s = \cos s. \end{aligned}$$

Hence, in this case, $c_Z(s)$ is a real (not complex) number for all $s \in R^1$. $\blacksquare$

Once we overcome our "fear" of imaginary and complex numbers, we can see that the characteristic function is actually much better in some ways than the moment-generating function. The main advantage is that, because $e^{isX} = \cos(sX) + i\sin(sX)$ and $|e^{isX}| = 1$, the characteristic function (unlike the moment-generating function) is always finite (although it could be a complex number).

Theorem 3.4.8 Let X be any random variable, and let s be any real number. Then $c_X(s)$ is finite.

The characteristic function has many properties similar to the moment-generating function. In particular, we have the following. (The proof is just like the proof of Theorem 3.4.3.)

Theorem 3.4.9 Let X be any random variable with its first k moments finite. Then $c_X(0) = 1$, $c'_X(0) = iE(X)$, $c''_X(0) = i^2E(X^2) = -E(X^2)$, etc. In general, $c_X^{(k)}(0) = i^kE(X^k)$, where $i = \sqrt{-1}$, and where $c_X^{(k)}$ is the kth derivative of c_X.

We also have the following. (The proof is just like the proof of Theorem 3.4.5.)

Theorem 3.4.10 Let X and Y be random variables which are *independent*. Then $c_{X+Y}(s) = c_X(s)\,c_Y(s)$.

For simplicity, we shall generally not use characteristic functions in this book. However, it is worth keeping in mind that whenever we do anything with moment-generating functions, we could usually do the same thing in greater generality using characteristic functions.

Summary of Section 3.4

- The probability-generating function of a random variable X is $r_X(t) = E(t^X)$.
- If X is discrete, then the derivatives of r_X satisfy $r_X^{(k)}(0) = k!\,P(X = k)$.
- The kth moment of a random variable X is $E(X^k)$.
- The moment-generating function of a random variable X is $m_X(s) = E(e^{sX}) = r_X(e^s)$.
- The derivatives of m_X satisfy $m_X^{(k)}(0) = E(X^k)$, for $k = 0, 1, 2, \ldots$.
- If X and Y are independent, then $r_{X+Y}(t) = r_X(t)\,r_Y(y)$ and $m_{X+Y}(s) = m_X(s)\,m_Y(s)$.
- If $m_X(s)$ is finite in a neighborhood of $s = 0$, then it uniquely characterizes the distribution of X.
- The characteristic function $c_X(s) = E(e^{isX})$ can be used in place of $m_X(s)$ to avoid infinities.

EXERCISES

3.4.1 Let Z be a discrete random variable with $P(Z = z) = 1/2^z$ for $z = 1, 2, 3, \ldots$.
(a) Compute $r_Z(t)$. Verify that $r_Z'(0) = P(Z = 1)$ and $r_Z''(0) = 2\,P(Z = 2)$.
(b) Compute $m_Z(t)$. Verify that $m_Z'(0) = E(Z)$ and $m_Z''(0) = E(Z^2)$.

3.4.2 Let $X \sim \text{Binomial}(n, \theta)$. Use m_X to prove that $\text{Var}(X) = n\theta(1 - \theta)$.

3.4.3 Let $Y \sim \text{Poisson}(\lambda)$. Use m_Y to compute the mean and variance of Y.

3.4.4 Let $Y = 3X + 4$. Compute $r_Y(t)$ in terms of r_X.

3.4.5 Let $Y = 3X + 4$. Compute $m_Y(s)$ in terms of m_X.

3.4.6 Let $X \sim \text{Binomial}(n, \theta)$. Compute $E(X^3)$, the third moment of X.

3.4.7 Let $Y \sim \text{Poisson}(\lambda)$. Compute $E(Y^3)$, the third moment of Y.

3.4.8 Suppose $P(X = 2) = 1/2$, $P(X = 5) = 1/3$, and $P(X = 7) = 1/6$.
(a) Compute $r_X(t)$ for $t \in R^1$.
(b) Verify that $r_X'(0) = P(X = 1)$ and $r_X''(0) = 2P(X = 2)$.
(c) Compute $m_X(s)$ for $s \in R^1$.
(d) Verify that $m_X'(0) = E(X)$ and $m_X''(0) = E(X^2)$.

PROBLEMS

3.4.9 Suppose $f_X(x) = 1/10$ for $0 < x < 10$, with $f_X(x) = 0$ otherwise.
(a) Compute $m_X(s)$ for $s \in R^1$.
(b) Verify that $m_X'(0) = E(X)$. (Hint: L'Hôpital's rule.)

3.4.10 Let $X \sim \text{Geometric}(\theta)$. Compute $r_X(t)$ and $r_X''(0)/2$.

3.4.11 Let $X \sim \text{Negative-Binomial}(r, \theta)$. Compute $r_X(t)$ and $r_X''(0)/2$.

3.4.12 Let $X \sim \text{Geometric}(\theta)$.
(a) Compute $m_X(s)$.
(b) Use m_X to compute the mean of X.
(c) Use m_X to compute the variance of X.

3.4.13 Let $X \sim$ Negative-Binomial(r, θ).
(a) Compute $m_X(s)$.
(b) Use m_X to compute the mean of X.
(c) Use m_X to compute the variance of X.

3.4.14 If $Y = a + bX$, where a and b are constants, then show that $r_Y(t) = t^a r_X(t^b)$ and $m_Y(t) = e^{at} m_X(bt)$.

3.4.15 Let $Z \sim N(\mu, \sigma^2)$. Show that

$$m_Z(s) = \exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}.$$

(Hint: Write $Z = \mu + \sigma X$ where $X \sim N(0, 1)$, and use Theorem 3.4.4.)

3.4.16 Let Y be distributed according to the Laplace distribution (see Problem 2.4.22).
(a) Compute $m_Y(s)$. (Hint: Break up the integral into two pieces.)
(b) Use m_Y to compute the mean of Y.
(c) Use m_Y to compute the variance of Y.

3.4.17 Compute the kth moment of the Weibull(α) distribution in terms of Γ (see Problem 2.4.19).

3.4.18 Compute the kth moment of the Pareto(α) distribution (see Problem 2.4.20). (Hint: Make the transformation $u = (1 + x)^{-1}$ and recall the beta distribution.)

3.4.19 Compute the kth moment of the Log-normal(τ) distribution (see Problem 2.6.17). (Hint: Make the transformation $z = \ln x$ and use Problem 3.4.15.)

3.4.20 Prove that the moment-generating function of the Gamma(α, λ) distribution is given by $\lambda^\alpha / (\lambda - t)^\alpha$ when $t < \lambda$.

3.4.21 Suppose that $X_i \sim$ Poisson(λ_i) and $X_1, \ldots, X_n$ are independent. Using moment-generating functions, determine the distribution of $Y = \sum_{i=1}^n X_i$.

3.4.22 Suppose that $X_i \sim$ Negative-Binomial(r_i, θ) and $X_1, \ldots, X_n$ are independent. Using moment-generating functions, determine the distribution of $Y = \sum_{i=1}^n X_i$.

3.4.23 Suppose that $X_i \sim$ Gamma(α_i, λ) and $X_1, \ldots, X_n$ are independent. Using moment-generating functions, determine the distribution of $Y = \sum_{i=1}^n X_i$.

3.4.24 Suppose $X_1, X_2, \ldots$ is i.i.d. Exponential(λ) and $N \sim$ Poisson(λ) independent of the $\{X_i\}$. Determine the moment-generating function of S_N. Determine the first moment of this distribution by differentiating this function.

3.4.25 Suppose $X_1, X_2, \ldots$ are i.i.d. Exponential(λ) random variables and $N \sim$ Geometric(θ), independent of the $\{X_i\}$. Determine the moment-generating function of S_N. Determine the first moment of this distribution by differentiating this function.

3.4.26 Let $X \sim$ Bernoulli(θ). Use $c_X(s)$ to compute the mean of X.

3.4.27 Let $Y \sim$ Binomial(n, θ).
(a) Compute the characteristic function $c_Y(s)$. (Hint: Make use of $c_X(s)$ in Problem 3.4.26.)
(b) Use $c_Y(s)$ to compute the mean of Y.

3.4.28 The characteristic function of the Cauchy distribution (see Problem 2.4.21) is given by $c(t) = e^{-|t|}$. Use this to determine the characteristic function of the sample

mean

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

based on a sample of n from the Cauchy distribution. Explain why this implies that the sample mean is also Cauchy distributed. What do you find surprising about this result?

3.4.29 The *kth cumulant* (when it exists) of a random variable X is obtained by calculating the kth derivative of $\ln c_X(s)$ with respect to s, evaluating this at $s = 0$, and dividing by i^k. Evaluate $c_X(s)$ and all the cumulants of the $N(\mu, \sigma^2)$ distribution.

3.5 Conditional Expectation

We have seen in Sections 1.5 and 2.8 that *conditioning* on some event, or some random variable, can change various probabilities. Now, because expectations are defined in terms of probabilities, it seems reasonable that expectations should also change when conditioning on some event or random variable. Such modified expectations are called *conditional expectations,* as we now discuss.

3.5.1 Discrete Case

The simplest case is when X is a discrete random variable, and A is some event of positive probability. We have the following.

Definition 3.5.1 Let X be a discrete random variable, and let A be some event with $P(A) > 0$. Then the *conditional expectation* of X, given A, is equal to

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

EXAMPLE 3.5.1

Consider rolling a fair six-sided die, so that $S = \{1, 2, 3, 4, 5, 6\}$. Let X be the number showing, so that $X(s) = s$ for $s \in S$. Let $A = \{3, 5, 6\}$ be the event that the die shows 3, 5, or 6. What is $E(X \mid A)$?

Here we know that

$$P(X = 3 \mid A) = P(X = 3 \mid X = 3, 5, \text{ or } 6) = 1/3$$

and that, similarly, $P(X = 5 \mid A) = P(X = 6 \mid A) = 1/3$. Hence,

$$\begin{aligned} E(X \mid A) &= \sum_{x \in R^1} x\, P(X = x \mid A) \\ &= 3\, P(X = 3 \mid A) + 5\, P(X = 5 \mid A) + 6\, P(X = 6 \mid A) \\ &= 3(1/3) + 5(1/3) + 6(1/3) = 14/3. \blacksquare \end{aligned}$$

Often we wish to condition on the value of some other random variable. If the other random variable is also discrete, and if the conditioned value has positive probability, then this works as above.

Definition 3.5.2 Let X and Y be discrete random variables, with $P(Y = y) > 0$. Then the *conditional expectation* of X, given $Y = y$, is equal to

$$E(X \mid Y = y) = \sum_{x \in R^1} x P(X = x \mid Y = y) = \sum_{x \in R^1} x \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

EXAMPLE 3.5.2
Suppose the joint probability function of X and Y is given by

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

Then

$$\begin{aligned} E(X \mid Y = 0) &= \sum_{x \in R^1} x\, P(X = x \mid Y = 0) \\ &= 5P(X = 5 \mid Y = 0) + 8P(X = 8 \mid Y = 0) \\ &= 5\frac{P(X = 5,\ Y = 0)}{P(Y = 0)} + 8\,\frac{P(X = 8,\ Y = 0)}{P(Y = 0)} \\ &= 5\frac{1/7}{1/7 + 3/7} + 8\,\frac{3/7}{1/7 + 3/7} = \frac{29}{4}. \end{aligned}$$

Similarly,

$$\begin{aligned} E(X \mid Y = 4) &= \sum_{x \in R^1} x P(X = x \mid Y = 4) \\ &= 5P(X = 5 \mid Y = 4) + 8P(X = 8 \mid Y = 4) \\ &= 5\frac{1/7}{1/7 + 1/7} + 8\frac{1/7}{1/7 + 1/7} = 13/2\,. \end{aligned}$$

Also,

$$\begin{aligned} E(X \mid Y = 3) &= \sum_{x \in R^1} x\, P(X = x \mid Y = 3) = 5P(X = 5 \mid Y = 3) \\ &= 5\frac{1/7}{1/7} = 5. \blacksquare \end{aligned}$$

Sometimes we wish to condition on a random variable Y, without specifying in advance on what value of Y we are conditioning. In this case, the conditional expectation $E(X \mid Y)$ is itself a random variable — namely, it depends on the (random) value of Y that occurs.

Definition 3.5.3 Let X and Y be discrete random variables. Then the *conditional expectation* of X, given Y, is the random variable $E(X \mid Y)$, which is equal to $E(X \mid Y = y)$ when $Y = y$. In particular, $E(X \mid Y)$ is a random variable that depends on the random value of Y.

EXAMPLE 3.5.3
Suppose again that the joint probability function of X and Y is given by

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

We have already computed that $E(X \mid Y = 0) = 29/4$, $E(X \mid Y = 4) = 13/2$, and $E(X \mid Y = 3) = 5$. We can express these results together by saying that

$$E(X \mid Y) = \begin{cases} 29/4 & Y = 0 \\ 5 & Y = 3 \\ 13/2 & Y = 4. \end{cases}$$

That is, $E(X \mid Y)$ is a random variable, which depends on the value of Y. Note that, because $P(Y = y) = 0$ for $y \neq 0, 3, 4$, the random variable $E(X \mid Y)$ is *undefined* in that case; but this is not a problem because that case will never occur. ∎

Finally, we note that just like for regular expectation, conditional expectation is linear.

Theorem 3.5.1 Let X_1, X_2, and Y be random variables; let A be an event; let a, b, and y be real numbers; and let $Z = aX_1 + bX_2$. Then
(a) $E(Z \mid A) = aE(X_1 \mid A) + bE(X_2 \mid A)$.
(b) $E(Z \mid Y = y) = aE(X_1 \mid Y = y) + bE(X_2 \mid Y = y)$.
(c) $E(Z \mid Y) = aE(X_1 \mid Y) + bE(X_2 \mid Y)$.

3.5.2 Absolutely Continuous Case

Suppose now that X and Y are jointly absolutely continuous. Then conditioning on $Y = y$, for some particular value of y, seems problematic, because $P(Y = y) = 0$. However, we have already seen in Section 2.8.2 that we can define a *conditional density* $f_{X|Y}(x|y)$ that gives us a density function for X, conditional on $Y = y$. And because density functions give rise to expectations, similarly conditional density functions give rise to conditional expectations, as follows.

Definition 3.5.4 Let X and Y be jointly absolutely continuous random variables, with joint density function $f_{X,Y}(x, y)$. Then the *conditional expectation* of X, given $Y = y$, is equal to

$$E(X \mid Y = y) = \int_{x \in R^1} x \, f_{X|Y}(x \mid y) \, dx = \int_{x \in R^1} x \, \frac{f_{X,Y}(x, y)}{f_Y(y)} \, dx.$$

EXAMPLE 3.5.4
Let X and Y be jointly absolutely continuous, with joint density function $f_{X,Y}$ given by

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

Then for $0 < y < 1$,

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

Hence,

$$\begin{aligned} E(X \mid Y = y) &= \int_{x \in R^1} x \frac{f_{X,Y}(x, y)}{f_Y(y)} \, dx = \int_0^1 x \, \frac{4x^2y + 2y^5}{4y/3 + 2y^5} \, dx \\ &= \frac{y + y^5}{4y/3 + 2y^5} = \frac{1 + y^4}{4/3 + 2y^4}. \blacksquare \end{aligned}$$

As in the discrete case, we often wish to condition on a random variable without specifying in advance the value of that variable. Thus, $E(X \mid Y)$ is again a random variable, depending on the random value of Y.

Definition 3.5.5 Let X and Y be jointly absolutely continuous random variables. Then the *conditional expectation* of X, given Y, is the random variable $E(X \mid Y)$, which is equal to $E(X \mid Y = y)$ when $Y = y$. Thus, $E(X \mid Y)$ is a random variable that depends on the random value of Y.

EXAMPLE 3.5.5
Let X and Y again have joint density

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

We already know that $E(X \mid Y = y) = \left(1 + y^4\right) / \left(4/3 + 2y^4\right)$. This formula is valid for any y between 0 and 1, so we conclude that $E(X \mid Y) = \left(1 + Y^4\right) / \left(4/3 + 2Y^4\right)$. Note that in this last formula, Y is a random variable, so $E(X \mid Y)$ is also a random variable. $\blacksquare$

Finally, we note that in the absolutely continuous case, conditional expectation is still linear, i.e., Theorem 3.5.1 continues to hold.

3.5.3 Double Expectations

Because the conditional expectation $E(X \mid Y)$ is itself a random variable (as a function of Y), it makes sense to take its expectation, $E\,(E(X \mid Y))$. This is a *double expectation*. One of the key results about conditional expectation is that it is always equal to $E(X)$.

Theorem 3.5.2 (*Theorem of total expectation*) If X and Y are random variables, then $E\,(E(X \mid Y)) = E(X)$.

This theorem follows as a special case of Theorem 3.5.3 on the next page. But it also makes sense intuitively. Indeed, conditioning on Y will change the conditional value of X in various ways, sometimes making it smaller and sometimes larger, depending on the value of Y. However, if we then average over all possible values of Y, these various effects will cancel out, and we will be left with just $E(X)$.

EXAMPLE 3.5.6
Suppose again that X and Y have joint probability function

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

Then we know that

$$E(X \mid Y = y) = \begin{cases} 29/4 & y = 0 \\ 5 & y = 3 \\ 13/2 & y = 4. \end{cases}$$

Also, $P(Y = 0) = 1/7 + 3/7 = 4/7$, $P(Y = 3) = 1/7$, and $P(Y = 4)$ $= 1/7 + 1/7 = 2/7$. Hence,

$$\begin{aligned} &E\,(E(X \mid Y)) \\ &= \sum_{y \in R^1} E(X \mid Y = y) P(Y = y) \\ &= E(X \mid Y = 0) P(Y = 0) + E(X \mid Y = 3) P(Y = 3) + E(X \mid Y = 4) P(Y = 4) \\ &= (29/4)(4/7) + (5)(1/7) + (13/2)(2/7) = 47/7. \end{aligned}$$

On the other hand, we compute directly that $E(X) = 5P(X = 5) + 8P(X = 8) =$ $5(3/7) + 8(4/7) = 47/7$. Hence, $E\,(E(X \mid Y)) = E(X)$, as claimed. ∎

EXAMPLE 3.5.7
Let X and Y again have joint density

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

We already know that

$$E(X \mid Y) = \left(1 + Y^4\right) / \left(4/3 + 2Y^4\right)$$

and that $f_Y(y) = 4y/3 + 2y^5$ for $0 \le y \le 1$. Hence,

$$\begin{aligned} &E\left(E(X \mid Y)\right) \\ &= E\left(\frac{1+Y^4}{4/3+2Y^4}\right) = \int_{-\infty}^{\infty} E(X \mid Y = y)\, f_Y(y)\, dy \\ &= \int_0^1 \frac{1+y^4}{4/3+2y^4}\,(4y/3+2y^5)\, dy = \int_0^1 (y+y^5)\, dy = 1/2 + 1/6 = 2/3. \end{aligned}$$

On the other hand,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f_{X,Y}(x,y)\, dy\, dx = \int_0^1\int_0^1 x\,(4x^2y + 2y^5)\, dy\, dx \\ &= \int_0^1 x\,(2x^2 + 2/6)\, dx = \int_0^1 (2x^3 + x/3)\, dx = 2/4 + 1/6 = 2/3\,. \end{aligned}$$

Hence, $E\left(E(X \mid Y)\right) = E(X)$, as claimed. ∎

Theorem 3.5.2 is a special case (with $g(y) \equiv 1$) of the following more general result, which in fact *characterizes* conditional expectation.

Theorem 3.5.3 Let X and Y be random variables, and let $g : R^1 \to R^1$ be any function. Then $E(g(Y)E(X \mid Y)) = E(g(Y)X)$.

PROOF See Section 3.8 for the proof of this result.

We also note the following related result. It says that, when conditioning on Y, any function of Y can be factored out since it is effectively a constant.

Theorem 3.5.4 Let X and Y be random variables, and let $g : R^1 \to R^1$ be any function. Then $E(g(Y)X \mid Y) = g(Y)E(X \mid Y)$.

PROOF See Section 3.8 for the proof of this result.

Finally, because conditioning *twice* on Y is the same as conditioning just once on Y, we immediately have the following.

Theorem 3.5.5 Let X and Y be random variables. Then $E(E(X \mid Y) \mid Y) = E(X \mid Y)$.

3.5.4 Conditional Variance (Advanced)

In addition to defining conditional expectation, we can define conditional variance. As usual, this involves the expected squared distance of a random variable to its mean.

However, in this case, the expectation is a conditional expectation. In addition, the mean is a conditional mean.

Definition 3.5.6 If X is a random variable, and A is an event with $P(A) > 0$, then the *conditional variance* of X, given A, is equal to

$$\text{Var}(X \mid A) = E((X - E(X \mid A))^2 \mid A) = E(X^2 \mid A) - (E(X \mid A))^2 .$$

Similarly, if Y is another random variable, then

$$\begin{aligned} \text{Var}(X \mid Y = y) &= E((X - E(X \mid Y = y))^2 \mid Y = y) \\ &= E(X^2 \mid Y = y) - (E(X \mid Y = y))^2 , \end{aligned}$$

and $\text{Var}(X \mid Y) = E((X - E(X \mid Y))^2 \mid Y) = E(X^2 \mid Y) - (E(X \mid Y))^2 .$

EXAMPLE 3.5.8
Consider again rolling a fair six-sided die, so that $S = \{1, 2, 3, 4, 5, 6\}$, with $P(s) = 1/6$ and $X(s) = s$ for $s \in S$, and with $A = \{3, 5, 6\}$. We have already computed that $P(X = s \mid A) = 1/3$ for $s \in A$, and that $E(X \mid A) = 14/3$. Hence,

$$\begin{aligned} &\text{Var}(X \mid A) = E\left((X - E(X \mid A))^2 \mid A\right) \\ &= E\left((X - 14/3)^2 \mid A\right) = \sum_{s \in S} (s - 14/3)^2 \, P(X = s \mid A) \\ &= (3 - 14/3)^2(1/3) + (5 - 14/3)^2(1/3) + (6 - 14/3)^2(1/3) = 14/9 \doteq 1.56. \end{aligned}$$

By contrast, because $E(X) = 7/2$, we have

$$\text{Var}(X) = E\left((X - E(X))^2\right) = \sum_{x=1}^{6} (x - 7/2)^2(1/6) = 35/12 \doteq 2.92.$$

Hence, we see that the conditional variance $\text{Var}(X \mid A)$ is much smaller than the unconditional variance $\text{Var}(X)$. This indicates that, in this example, once we know that event A has occurred, we know more about the value of X than we did originally. ∎

EXAMPLE 3.5.9
Suppose X and Y have joint density function

$$f_{X,Y}(x, y) = \begin{cases} 8xy & 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

We have $f_Y(y) = 4y^3$, $f_{X|Y}(x \mid y) = 8xy/4y^3 = 2x/y^2$ for $0 < x < y$ and so

$$E(X \mid Y = y) = \int_0^y x \frac{2x}{y^2} \, dx = \int_0^y \frac{2x^2}{y^2} \, dx = \frac{2y^3}{3y^2} = \frac{2y}{3}.$$

Therefore,

$$\begin{aligned} \text{Var}(X \mid Y = y) &= E\left((X - E(X \mid Y = y))^2 \mid Y = y\right) \\ &= \int_0^y \left(x - \frac{2y}{3}\right)^2 \frac{2x}{y^2} \, dx = \frac{1}{2y^2} - \frac{8}{9y} + \frac{4}{9}. \end{aligned}$$ ∎

Finally, we note that conditional expectation and conditional variance satisfy the following useful identity.

Theorem 3.5.6 For random variables X and Y,

$$\text{Var}(X) = \text{Var}\left(E(X \mid Y)\right) + E\left(\text{Var}(X \mid Y)\right).$$

PROOF See Section 3.8 for the proof of this result.

Summary of Section 3.5

- If X is discrete, then the conditional expectation of X, given an event A, is equal to $E(X \mid A) = \sum_{x \in R^1} x P(X = x \mid A)$.
- If X and Y are discrete random variables, then $E(X \mid Y)$ is itself a random variable, with $E(X \mid Y)$ equal to $E(X \mid Y = y)$ when $Y = y$.
- If X and Y are jointly absolutely continuous, then $E(X \mid Y)$ is itself a random variable, with $E(X \mid Y)$ equal to $E(X \mid Y = y)$ when $Y = y$, where $E(X \mid Y = y) = \int x\, f_{X|Y}(x \mid y)\, dx$.
- Conditional expectation is linear.
- We always have that $E\left(g(Y)\, E(X \mid Y)\right) = E\left(g(Y)\, X\right)$, and $E\left(E(X \mid Y) | Y\right) = E(X \mid Y)$.
- Conditional variance is given by $\text{Var}(X \mid Y) = E(X^2 \mid Y) - (E(X \mid Y))^2$.

EXERCISES

3.5.1 Suppose X and Y are discrete, with

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

(a) Compute $E(X \mid Y = 3)$.
(b) Compute $E(Y \mid X = 3)$.
(c) Compute $E(X \mid Y)$.
(d) Compute $E(Y \mid X)$.

3.5.2 Suppose X and Y are jointly absolutely continuous, with

$$f_{X,Y}(x, y) = \begin{cases} 9(xy + x^5 y^5)/16,000,900 & 0 \le x \le 4,\ 0 \le y \le 5 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $f_X(x)$.

(b) Compute $f_Y(y)$.
(c) Compute $E(X \mid Y)$.
(d) Compute $E(Y \mid X)$.
(e) Compute $E\,(E(X \mid Y))$, and verify that it is equal to $E(X)$.

3.5.3 Suppose X and Y are discrete, with

$$p_{X,Y}(x, y) = \begin{cases} 1/11 & x = -4,\ y = 2 \\ 2/11 & x = -4,\ y = 3 \\ 4/11 & x = -4,\ y = 7 \\ 1/11 & x = 6,\ y = 2 \\ 1/11 & x = 6,\ y = 3 \\ 1/11 & x = 6,\ y = 7 \\ 1/11 & x = 6,\ y = 13 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $E(Y \mid X = 6)$.
(b) Compute $E(Y \mid X = -4)$.
(c) Compute $E(Y \mid X)$.

3.5.4 Let $p_{X,Y}$ be as in the previous exercise.
(a) Compute $E(X \mid Y = 2)$.
(b) Compute $E(X \mid Y = 3)$.
(c) Compute $E(X \mid Y = 7)$.
(d) Compute $E(X \mid Y = 13)$.
(e) Compute $E(X \mid Y)$.

3.5.5 Suppose that a student must choose one of two summer job offers. If it is not necessary to take a summer course, then a job as a waiter will produce earnings (rounded to the nearest \$1000) with the following probability distribution.

\$1000	\$2000	\$3000	\$4000
0.1	0.3	0.4	0.2

If it is necessary to take a summer course, then a part-time job at a hotel will produce earnings (rounded to the nearest \$1000) with the following probability distribution.

\$1000	\$2000	\$3000	\$4000
0.3	0.4	0.2	0.1

If the probability that the student will have to take the summer course is 0.6, then determine the student's expected summer earnings.

3.5.6 Suppose you roll two fair six-sided dice. Let X be the number showing on the first die, and let Z be the sum of the two numbers showing.
(a) Compute $E(X)$.
(b) Compute $E(Z \mid X = 1)$.
(c) Compute $E(Z \mid X = 6)$.
(d) Compute $E(X \mid Z = 2)$.
(e) Compute $E(X \mid Z = 4)$.

(f) Compute $E(X \mid Z = 6)$.
(g) Compute $E(X \mid Z = 7)$.
(h) Compute $E(X \mid Z = 11)$.

3.5.7 Suppose you roll two fair six-sided dice. Let Z be the sum of the two numbers showing, and let W be the product of the two numbers showing.
(a) Compute $E(Z \mid W = 4)$.
(b) Compute $E(W \mid Z = 4)$.

3.5.8 Suppose you roll one fair six-sided die and then flip as many coins as the number showing on the die. (For example, if the die shows 4, then you flip four coins.) Let X be the number showing on the die, and Y be the number of heads obtained.
(a) Compute $E(Y \mid X = 5)$.
(b) Compute $E(X \mid Y = 0)$.
(c) Compute $E(X \mid Y = 2)$.

3.5.9 Suppose you flip three fair coins. Let X be the number of heads obtained, and let $Y = 1$ if the first coin shows heads, otherwise $Y = 0$.
(a) Compute $E(X \mid Y = 0)$.
(b) Compute $E(X \mid Y = 1)$.
(c) Compute $E(Y \mid X = 0)$.
(d) Compute $E(Y \mid X = 1)$.
(e) Compute $E(Y \mid X = 2)$.
(f) Compute $E(Y \mid X = 3)$.
(g) Compute $E(Y \mid X)$.
(h) Verify directly that $E[E(Y \mid X)] = E(Y)$.

3.5.10 Suppose you flip one fair coin and roll one fair six-sided die. Let X be the number showing on the die, and let $Y = 1$ if the coin is heads with $Y = 0$ if the coin is tails. Let $Z = XY$.
(a) Compute $E(Z)$.
(b) Compute $E(Z \mid X = 4)$.
(c) Compute $E(Y \mid X = 4)$.
(d) Compute $E(Y \mid Z = 4)$.
(e) Compute $E(X \mid Z = 4)$.

3.5.11 Suppose X and Y are jointly absolutely continuous, with joint density function $f_{X,Y}(x, y) = (6/19)(x^2 + y^3)$ for $0 < x < 2$ and $0 < y < 1$, otherwise $f_{X,Y}(x, y) = 0$.
(a) Compute $E(X)$.
(b) Compute $E(Y)$.
(c) Compute $E(X \mid Y)$.
(d) Compute $E(Y \mid X)$.
(e) Verify directly that $E[E(X \mid Y)] = E(X)$.
(f) Verify directly that $E[E(Y \mid X)] = E(Y)$.

PROBLEMS

3.5.12 Suppose there are two urns. Urn I contains 100 chips: 30 are labelled 1, 40 are labelled 2, and 30 are labelled 3. Urn 2 contains 100 chips: 20 are labelled 1,

50 are labelled 2, and 30 are labelled 3. A coin is tossed and if a head is observed, then a chip is randomly drawn from urn 1, otherwise a chip is randomly drawn from urn 2. The value Y on the chip is recorded. If an occurrence of a head on the coin is denoted by $X = 1$, a tail by $X = 0$, and $X \sim \text{Bernoulli}(3/4)$, then determine $E(X \mid Y)$, $E(Y \mid X)$, $E(Y)$, and $E(X)$.

3.5.13 Suppose that five coins are each tossed until the first head is obtained on each coin and where each coin has probability θ of producing a head. If you are told that the total number of tails observed is $Y = 10$, then determine the expected number of tails observed on the first coin.

3.5.14 *(Simpson's paradox)* Suppose that the conditional distributions of Y, given X, are shown in the following table. For example, $p_{Y|X}(1 \mid i)$ could correspond to the probability that a randomly selected heart patient at hospital i has a successful treatment.

$p_{Y\|X}(0 \mid 1)$ 0.030	$p_{Y\|X}(1 \mid 1)$ 0.970
$p_{Y\|X}(0 \mid 2)$ 0.020	$p_{Y\|X}(1 \mid 2)$ 0.980

(a) Compute $E(Y \mid X)$.

(b) Now suppose that patients are additionally classified as being seriously ill ($Z = 1$), or not seriously ill ($Z = 0$). The conditional distributions of Y, given (X, Z), are shown in the following tables. Compute $E(Y \mid X, Z)$.

$p_{Y\|X,Z}(0 \mid 1, 0)$ 0.010	$p_{Y\|X,Z}(1 \mid 1, 0)$ 0.990
$p_{Y\|X,Z}(0 \mid 2, 0)$ 0.013	$p_{Y\|X,Z}(1 \mid 2, 0)$ 0.987

$p_{Y\|X,Z}(0 \mid 1, 1)$ 0.038	$p_{Y\|X,Z}(1 \mid 1, 1)$ 0.962
$p_{Y\|X,Z}(0 \mid 2, 1)$ 0.040	$p_{Y\|X,Z}(1 \mid 2, 1)$ 0.960

(c) Explain why the conditional distributions in part (a) indicate that hospital 2 is the better hospital for a patient who needs to undergo this treatment, but all the conditional distributions in part (b) indicate that hospital 1 is the better hospital. This phenomenon is known as Simpson's paradox.

(d) Prove that, in general, $p_{Y|X}(y \mid x) = \sum_z p_{Y|X,Z}(y \mid x, z)\, p_{Z|X}(z \mid x)$ and $E(Y \mid X) = E(E(Y \mid X, Z) \mid X)$.

(e) If the conditional distributions $p_{Z|X}(\cdot \mid x)$, corresponding to the example discussed in parts (a) through (c) are given in the following table, verify the result in part (d) numerically and explain how this resolves Simpson's paradox.

$p_{Z\|X}(0 \mid 1)$ 0.286	$p_{Z\|X}(1 \mid 1)$ 0.714
$p_{Z\|X}(0 \mid 2)$ 0.750	$p_{Z\|X}(1 \mid 2)$ 0.250

3.5.15 Present an example of a random variable X, and an event A with $P(A) > 0$, such that $\text{Var}(X \mid A) > \text{Var}(X)$. (Hint: Suppose $S = \{1, 2, 3\}$ with $X(s) = s$, and $A = \{1, 3\}$.)

3.5.16 Suppose that X, given $Y = y$, is distributed Gamma(α, y) and that the marginal distribution of Y is given by $1/Y \sim$ Exponential(λ). Determine $E(X)$.

3.5.17 Suppose that $(X, Y) \sim$ Bivariate Normal $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Use (2.7.1) (when given $Y = y$) and its analog (when given $X = x$) to determine $E(X \mid Y)$, $E(Y \mid X)$, $\text{Var}(X \mid Y)$, and $\text{Var}(Y \mid X)$.

3.5.18 Suppose that $(X_1, X_2, X_3) \sim$ Multinomial$(n, \theta_1, \theta_2, \theta_3)$. Determine $E(X_1 \mid X_2)$ and $\text{Var}(X_1 \mid X_2)$. (Hint: Show that X_1, given $X_2 = x_2$, has a binomial distribution.)

3.5.19 Suppose that $(X_1, X_2) \sim$ Dirichlet$(\alpha_1, \alpha_2, \alpha_3)$. Determine $E(X_1 \mid X_2)$ and $\text{Var}(X_1 \mid X_2)$. (Hint: First show that $X_1/(1 - x_2)$, given $X_2 = x_2$, has a beta distribution and then use Problem 3.3.24.)

3.5.20 Let $f_{X,Y}$ be as in Exercise 3.5.2.
(a) Compute $\text{Var}(X)$.
(b) Compute $\text{Var}(E(X \mid Y))$.
(c) Compute $\text{Var}(X \mid Y)$.
(d) Verify that $\text{Var}(X) = \text{Var}(E(X \mid Y)) + E(\text{Var}(X \mid Y))$.

3.5.21 Suppose we have three discrete random variables X, Y, and Z. We say that X and Y are *conditionally independent*, given Z, if

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

for every x, y, and z such that $P(Z = z) > 0$. Prove that when X and Y are conditionally independent, given Z, then

$$E(g(X)h(Y) \mid Z) = E(g(X) \mid Z)\, E(h(Y) \mid Z).$$

3.6 Inequalities

Expectation and variance are closely related to the underlying distributions of random variables. This relationship allows us to prove certain inequalities that are often very useful. We begin with a classic result, Markov's inequality, which is very simple but also very useful and powerful.

Theorem 3.6.1 (*Markov's inequality*) If X is a nonnegative random variable, then for all $a > 0$,

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

That is, the probability that X exceeds any given value a is no more than the mean of X divided by a.

PROOF Define a new random variable Z by

$$Z = \begin{cases} a & X \geq a \\ 0 & X < a. \end{cases}$$

Then clearly $Z \leq X$, so that $E(Z) \leq E(X)$ by monotonicity. On the other hand,

$$E(Z) = a\,P(Z = a) + 0\,P(Z = 0) = a\,P(Z = a) = a\,P(X \geq a).$$

So, $E(X) \geq E(Z) = a\,P(X \geq a)$. Rearranging, $P(X \geq a) \leq E(X)/a$, as claimed. ∎

Intuitively, Markov's inequality says that if the expected value of X is small, then it is unlikely that X will be too large. We now consider some applications of Theorem 3.6.1.

EXAMPLE 3.6.1
Suppose $P(X = 3) = 1/2$, $P(X = 4) = 1/3$, and $P(X = 7) = 1/6$. Then $E(X) = 3(1/2) + 4(1/3) + 7(1/6) = 4$. Hence, setting $a = 6$, Markov's inequality says that $P(X \geq 6) \leq 4/6 = 2/3$. In fact, $P(X \geq 6) = 1/6 < 2/3$. ∎

EXAMPLE 3.6.2
Suppose $P(X = 2) = P(X = 8) = 1/2$. Then $E(X) = 2(1/2) + 8(1/2) = 5$. Hence, setting $a = 8$, Markov's inequality says that $P(X \geq 8) \leq 5/8$. In fact, $P(X \geq 8) = 1/2 < 5/8$. ∎

EXAMPLE 3.6.3
Suppose $P(X = 0) = P(X = 2) = 1/2$. Then $E(X) = 0(1/2) + 2(1/2) = 1$. Hence, setting $a = 2$, Markov's inequality says that $P(X \geq 2) \leq 1/2$. In fact, $P(X \geq 2) = 1/2$, so Markov's inequality is an *equality* in this case. ∎

Markov's inequality is also used to prove Chebychev's inequality, perhaps the most important inequality in all of probability theory.

Theorem 3.6.2 (*Chebychev's inequality*) Let Y be an arbitrary random variable, with finite mean μ_Y. Then for all $a > 0$,

$$P\left(|Y - \mu_Y| \geq a\right) \leq \frac{\text{Var}(Y)}{a^2}.$$

PROOF Set $X = (Y - \mu_Y)^2$. Then X is a nonnegative random variable. Thus, using Theorem 3.6.1, we have $P\left(|Y - \mu_Y| \geq a\right) = P\left(X \geq a^2\right) \leq E(X)/a^2 = \text{Var}(Y)/a^2$, and this establishes the result. ∎

Intuitively, Chebychev's inequality says that if the variance of Y is small, then it is unlikely that Y will be too far from its mean value μ_Y. We now consider some examples.

EXAMPLE 3.6.4
Suppose again that $P(X = 3) = 1/2$, $P(X = 4) = 1/3$, and $P(X = 7) = 1/6$. Then $E(X) = 4$, as above. Also, $E(X^2) = 9(1/2) + 16(1/3) + 49(1/6) = 18$, so that $\text{Var}(X) = 18 - 4^2 = 2$. Hence, setting $a = 1$, Chebychev's inequality says that $P(|X - 4| \geq 1) \leq 2/1^2 = 2$, which tells us nothing because we always have $P(|X - 4| \geq 1) \leq 1$. On the other hand, setting $a = 3$, we get $P(|X - 4| \geq 3) \leq 2/3^2 = 2/9$, which is true because in fact $P(|X - 4| \geq 3) = P(X = 7) = 1/6 < 2/9$. ∎

EXAMPLE 3.6.5
Let $X \sim \text{Exponential}(3)$, and let $a = 5$. Then $E(X) = 1/3$ and $\text{Var}(X) = 1/9$. Hence, by Chebychev's inequality with $a = 1/2$, $P(|X - 1/3| \geq 1/2) \leq (1/9)/(1/2)^2 = 4/9$. On the other hand, because $X \geq 0$, $P(|X - 1/3| \geq 1/2) = P(X \geq 5/6)$, and by Markov's inequality, $P(X \geq 5/6) \leq (1/3)/(5/6) = 2/5$. Because $2/5 < 4/9$, we actually get a better bound from Markov's inequality than from Chebychev's inequality in this case. ∎

EXAMPLE 3.6.6
Let $Z \sim N(0, 1)$, and $a = 5$. Then by Chebychev's inequality, $P(|Z| \geq 5) \leq 1/5$. ∎

EXAMPLE 3.6.7
Let X be a random variable having *very small* variance. Then Chebychev's inequality says that $P(|X - \mu_X| \geq a)$ is small whenever a is not too small. In other words, usually $|X - \mu_X|$ is very small, i.e., $X \approx \mu_X$. This makes sense, because if the variance of X is very small, then usually X is very close to its mean value μ_X. ∎

Inequalities are also useful for covariances, as follows.

Theorem 3.6.3 (*Cauchy–Schwartz inequality*) Let X and Y be arbitrary random variables, each having finite, nonzero variance. Then

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\,\text{Var}(Y)}.$$

Furthermore, if $\text{Var}(Y) > 0$, then equality is attained if and only if $X - \mu_X = \lambda(Y - \mu_Y)$ where $\lambda = \text{Cov}(X, Y)/\text{Var}(Y)$.

PROOF See Section 3.8 for the proof. ∎

The Cauchy–Schwartz inequality says that if the variance of X or Y is small, then the covariance of X and Y must also be small.

EXAMPLE 3.6.8
Suppose $X = C$ is a constant. Then $\text{Var}(X) = 0$. It follows from the Cauchy–Schwartz inequality that, for *any* random variable Y, we must have $\text{Cov}(X, Y) \leq (\text{Var}(X)\,\text{Var}(Y))^{1/2} = (0\,\text{Var}(Y))^{1/2} = 0$, so that $\text{Cov}(X, Y) = 0$. ∎

Recalling that the *correlation* of X and Y is defined by

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}},$$

we immediately obtain the following important result (which has already been referred to, back when correlation was first introduced).

Corollary 3.6.1 Let X and Y be arbitrary random variables, having finite means and finite, nonzero variances. Then $|\text{Corr}(X, Y)| \leq 1$. Furthermore, $|\text{Corr}(X, Y)| = 1$ if and only if

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

So the correlation between two random variables is always between -1 and 1. We also see that X and Y are linearly related if and only if $|\text{Corr}(X, Y)| = 1$, and that this relationship is increasing (positive slope) when $\text{Corr}(X, Y) = 1$ and decreasing (negative slope) when $\text{Corr}(X, Y) = -1$.

3.6.1 Jensen's Inequality (Advanced)

Finally, we develop a more advanced inequality that is sometimes very useful. A function f is called *convex* if for every $x < y$, the line segment from $(x, f(x))$ to $(y, f(y))$ lies entirely *above* the graph of f, as depicted in Figure 3.6.1.

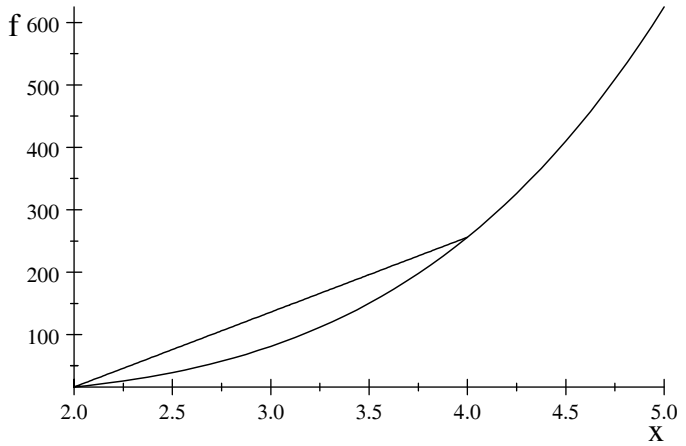

Figure 3.6.1: Plot of the convex function $f(x) = x^4$ and the line segment joining $(2, f(2))$ to $(4, f(4))$.

In symbols, we require that for every $x < y$ and every $0 < \lambda < 1$, we have $\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y)$. Examples of convex functions include $f(x) = x^2$, $f(x) = x^4$, and $f(x) = \max(x, C)$ for any real number C. We have the following.

Theorem 3.6.4 (*Jensen's inequality*) Let X be an arbitrary random variable, and let $f : R^1 \to R^1$ be a convex function such that $E(f(X))$ is finite. Then $f(E(X)) \leq E(f(X))$. Equality occurs if and only if $f(X) = a + bX$ for some a and b.

PROOF Because f is convex, we can find a linear function $g(x) = ax + b$ such that $g(E(X)) = f(E(X))$ and $g(x) \leq f(x)$ for all $x \in R^1$ (see, for example, Figure 3.6.2).

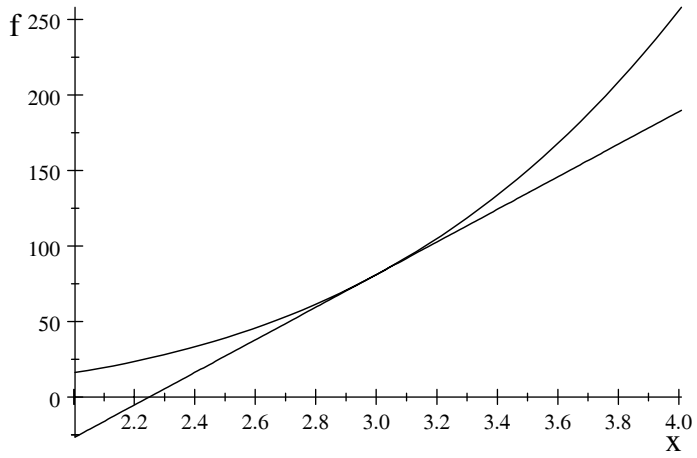

Figure 3.6.2: Plot of the convex function $f(x) = x^4$ and the function $g(x) = 81 + 108(x - 3)$, satisfying $g(x) \leq f(x)$ on the interval $(2, 4)$.

But then using monotonicity and linearity, we have $E(f(X)) \geq E(g(X)) = E(aX + b) = aE(X) + b = g(E(X)) = f(E(X))$, as claimed.

We have equality if and only if $0 = E(f(X) - g(X))$. Because $f(X) - g(X) \geq 0$, this occurs (using Challenge 3.3.29) if and only if $f(X) = g(X) = aX + b$ with probability 1. ∎

EXAMPLE 3.6.9
Let X be a random variable with finite variance. Then setting $f(x) = x^2$, Jensen's inequality says that $E(X^2) \geq (E(X))^2$. Of course, we already knew this because $E(X^2) - (E(X))^2 = \text{Var}(X) \geq 0$. ∎

EXAMPLE 3.6.10
Let X be a random variable with finite fourth moment. Then setting $f(x) = x^4$, Jensen's inequality says that $E(X^4) \geq (E(X))^4$. ∎

EXAMPLE 3.6.11
Let X be a random variable with finite mean, and let $M \in R^1$. Then setting $f(x) = \max(x, M)$, we have that $E(\max(X, M)) \geq \max(E(X), M)$ by Jensen's inequality. In fact, we could also have deduced this from the monotonicity property of expectation, using the two inequalities $\max(X, M) \geq X$ and $\max(X, M) \geq M$. ∎

Summary of Section 3.6

- For nonnegative X, Markov's inequality says $P(X \geq a) \leq E(X)/a$.
- Chebychev's inequality says $P(|Y - \mu_Y| \geq a) \leq \text{Var}(Y)/a^2$.
- The Cauchy–Schwartz inequality says $|\text{Cov}(X, Y)| \leq (\text{Var}(X)\,\text{Var}(Y))^{1/2}$, so that $|\text{Corr}(X, Y)| \leq 1$.

- Jensen's inequality says $f(E(X)) \leq E(f(X))$ whenever f is *convex*.

EXERCISES

3.6.1 Let $Z \sim$ Poisson(3). Use Markov's inequality to get an upper bound on $P(Z \geq 7)$.

3.6.2 Let $X \sim$ Exponential(5). Use Markov's inequality to get an upper bound on $P(X \geq 3)$ and compare it with the precise value.

3.6.3 Let $X \sim$ Geometric$(1/2)$.
(a) Use Markov's inequality to get an upper bound on $P(X \geq 9)$.
(b) Use Markov's inequality to get an upper bound on $P(X \geq 2)$.
(c) Use Chebychev's inequality to get an upper bound on $P(|X - 1| \geq 1)$.
(d) Compare the answers obtained in parts (b) and (c).

3.6.4 Let $Z \sim N(5, 9)$. Use Chebychev's inequality to get an upper bound on $P(|Z - 5| \geq 30)$.

3.6.5 Let $W \sim$ Binomial$(100, 1/2)$, as in the number of heads when flipping 100 fair coins. Use Chebychev's inequality to get an upper bound on $P(|W - 50| \geq 10)$.

3.6.6 Let $Y \sim N(0, 100)$, and let $Z \sim$ Binomial$(80, 1/4)$. Determine (with explanation) the largest and smallest possible values of $\text{Cov}(Y, Z)$.

3.6.7 Let $X \sim$ Geometric$(1/11)$. Use Jensen's inequality to determine a lower bound on $E(X^4)$, in two different ways.
(a) Apply Jensen's inequality to X with $f(x) = x^4$.
(b) Apply Jensen's inequality to X^2 with $f(x) = x^2$.

3.6.8 Let X be the number showing on a fair six-sided die. What bound does Chebychev's inequality give for $P(X \geq 5 \text{ or } X \leq 2)$?

3.6.9 Suppose you flip four fair coins. Let Y be the number of heads obtained.
(a) What bound does Chebychev's inequality give for $P(Y \geq 3 \text{ or } Y \leq 1)$?
(b) What bound does Chebychev's inequality give for $P(Y \geq 4 \text{ or } Y \leq 0)$?

3.6.10 Suppose W has density function $f(w) = 3w^2$ for $0 < w < 1$, otherwise $f(w) = 0$.
(a) Compute $E(W)$.
(b) What bound does Chebychev's inequality give for $P(|W - E(W)| \geq 1/4)$?

3.6.11 Suppose Z has density function $f(z) = z^3/4$ for $0 < z < 2$, otherwise $f(z) = 0$.
(a) Compute $E(Z)$.
(b) What bound does Chebychev's inequality give for $P(|Z - E(Z)| \geq 1/2)$?

3.6.12 Suppose $\text{Var}(X) = 4$ and $\text{Var}(Y) = 9$.
(a) What is the largest possible value of $\text{Cov}(X, Y)$?
(b) What is the smallest possible value of $\text{Cov}(X, Y)$?
(c) Suppose $Z = 3X/2$. Compute $\text{Var}(Z)$ and $\text{Cov}(X, Z)$, and compare your answer with part (a).
(d) Suppose $W = -3X/2$. Compute $\text{Var}(W)$ and $\text{Cov}(W, Z)$, and compare your answer with part (b).

3.6.13 Suppose a species of beetle has length 35 millimeters on average. Find an upper bound on the probability that a randomly chosen beetle of this species will be over 80 millimeters long.

PROBLEMS

3.6.14 Prove that for any $\epsilon > 0$ and $\delta > 0$, there is a positive integer M, such that if X is the number of heads when flipping M fair coins, then $P(|(X/M) - (1/2)| \geq \delta) \leq \epsilon$.

3.6.15 Prove that for any μ and $\sigma^2 > 0$, there is $a > 0$ and a random variable X with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$, such that Chebychev's inequality holds with equality, i.e., such that $P(|X - \mu| \geq a) = \sigma^2/a^2$.

3.6.16 Suppose that (X, Y) is uniform on the set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where the $x_1, \ldots, x_n$ are distinct values and the $y_1, \ldots, y_n$ are distinct values.

(a) Prove that X is uniformly distributed on $x_1, \ldots, x_n$, with mean given by $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ and variance given by $\hat{s}_X^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

(b) Prove that the correlation coefficient between X and Y is given by

$$r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\hat{s}_{XY}}{\hat{s}_X \hat{s}_Y}$$

where $\hat{s}_{XY} = n^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. The value $\hat{s}_{XY}$ is referred to as the *sample covariance* and r_{XY} is referred to as the *sample correlation coefficient* when the values $(x_1, y_1), \ldots, (x_n, y_n)$ are an observed sample from some bivariate distribution.

(c) Argue that r_{XY} is also the correlation coefficient between X and Y when we drop the assumption of distinctness for the x_i and y_i.

(d) Prove that $-1 \leq r_{XY} \leq 1$ and state the conditions under which $r_{XY} = \pm 1$.

3.6.17 Suppose that X is uniformly distributed on $\{x_1, \ldots, x_n\}$ and so has mean $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ and variance $\hat{s}_X^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ (see Problem 3.6.16(a)). What is the largest proportion of the values x_i that can lie outside $\left(\bar{x} - 2\hat{s}_X, \bar{x} + 2\hat{s}_X\right)$?

3.6.18 Suppose that X is distributed with density given by $f_X(x) = 2/x^3$ for $x > 1$ and is 0 otherwise.

(a) Prove that f_X is a density.

(b) Calculate the mean of X.

(c) Compute $P(X \geq k)$ and compare this with the upper bound on this quantity given by Markov's inequality.

(d) What does Chebyshev's inequality say in this case?

3.6.19 Let $g(x) = \max(-x, -10)$.

(a) Verify that g is a convex function.

(b) Suppose $Z \sim$ Exponential(5). Use Jensen's inequality to obtain a lower bound on $E(g(Z))$.

3.6.20 It can be shown that a function f, with continuous second derivative, is convex on (a, b) if $f''(x) > 0$ for all $x \in (a, b)$.

(a) Use the above fact to show that $f(x) = x^p$ is convex on $(0, \infty)$ whenever $p \geq 1$.

(b) Use part (a) to prove that $\left(E\left(|X|^p\right)\right)^{1/p} \geq |E(X)|$ whenever $p \geq 1$.
(c) Prove that $\text{Var}(X) = 0$ if and only if X is degenerate at a constant c.

CHALLENGES

3.6.21 Determine (with proof) all functions that are convex and whose *negatives* are also convex. (That is, find all functions f such that f is convex, and also $-f$ is convex.)

3.7 General Expectations (Advanced)

So far we have considered expected values separately for discrete and absolutely continuous random variables only. However, this separation into two different "cases" may seem unnatural. Furthermore, we know that some random variables are neither discrete nor continuous — for example, mixtures of discrete and continuous distributions.

Hence, it seems desirable to have a more general definition of expected value. Such generality is normally considered in the context of general measure theory, an advanced mathematical subject. However, it is also possible to give a general definition in elementary terms, as follows.

Definition 3.7.1 Let X be an arbitrary random variable (perhaps neither discrete nor continuous). Then the *expected value* of X is given by

$$E(X) = \int_0^\infty P(X > t)\,dt - \int_{-\infty}^0 P(X < t)\,dt,$$

provided either $\int_0^\infty P(X > t)\,dt < \infty$ or $\int_{-\infty}^0 P(X < t)\,dt < \infty$.

This definition appears to contradict our previous definitions of $E(X)$. However, in fact, there is no contradiction, as the following theorem shows.

Theorem 3.7.1
(a) Let X be a discrete random variable with distinct possible values $x_1, x_2, \ldots$, and put $p_i = P(X = x_i)$. Then Definition 3.7.1 agrees with the previous definition of $E(X)$. That is,

$$\int_0^\infty P(X > t)\,dt - \int_{-\infty}^0 P(X < t)\,dt = \sum_i x_i\, p_i.$$

(b) Let X be an absolutely continuous random variable with density f_X. Then Definition 3.7.1 agrees with the previous definition of $E(X)$. That is,

$$\int_0^\infty P(X > t)\,dt - \int_{-\infty}^0 P(X < t)\,dt = \int_{-\infty}^\infty x f_X(x)\,dx.$$

PROOF The key to the proof is switching the order of the integration/summation.

(a) We have

$$\int_0^\infty P(X > t)\,dt = \int_0^\infty \sum_{i,x_i>t} p_i\,dt = \sum_i p_i \int_0^{x_i} dt = \sum_i p_i\,x_i,$$

as claimed.

(b) We have

$$\begin{aligned}\int_0^\infty P(X > t)\,dt &= \int_0^\infty \left(\int_t^\infty f_X(x)\,dx\right) dt = \int_0^\infty \left(\int_0^x f_X(x)\,dt\right) dx \\ &= \int_0^\infty x\,f_X(x)\,dx.\end{aligned}$$

Similarly,

$$\begin{aligned}\int_{-\infty}^0 P(X < t)\,dt &= \int_{-\infty}^0 \left(\int_{-\infty}^t f_X(x)\,dx\right) dt = \int_{-\infty}^0 \left(\int_x^0 f_X(x)\,dt\right) dx \\ &= \int_{-\infty}^0 (-x)\,f_X(x)\,dx.\end{aligned}$$

Hence,

$$\begin{aligned}\int_0^\infty P(X > t)\,dt - \int_{-\infty}^0 P(X < t)\,dt &= \int_0^\infty x\,f_X(x)\,dx - \int_{-\infty}^0 (-x)\,f_X(x)\,dx \\ &= \int_{-\infty}^\infty x\,f_X(x)\,dx,\end{aligned}$$

as claimed. ∎

In other words, Theorem 3.7.1 says that Definition 3.7.1 includes our previous definitions of expected value, for both discrete and absolutely continuous random variables, while working for *any* random variable at all. (Note that to apply Definition 3.7.1 we take an *integral*, not a *sum*, regardless of whether X is discrete or continuous!)

Furthermore, Definition 3.7.1 preserves the key properties of expected value, as the following theorem shows. (We omit the proof here, but see Challenge 3.7.10 for a proof of part (c).)

Theorem 3.7.2 Let X and Y be arbitrary random variables, perhaps neither discrete nor continuous, with expected values defined by Definition 3.7.1.
(a) (*Linearity*) If a and b are any real numbers, then $E(aX+bY) = aE(X)+bE(Y)$.
(b) If X and Y are independent, then $E(XY) = E(X)\,E(Y)$.
(c) (*Monotonicity*) If $X \le Y$, then $E(X) \le E(Y)$.

Definition 3.7.1 also tells us about expected values of mixture distributions, as follows.

Theorem 3.7.3 For $1 \leq i \leq k$, let Y_i be a random variable with cdf F_i. Let X be a random variable whose cdf corresponds to a finite mixture (as in Section 2.5.4) of the cdfs of the Y_i, so that $F_X(x) = \sum_i p_i F_i(x)$, where $p_i \geq 0$ and $\sum_i p_i = 1$. Then $E(X) = \sum_i p_i E(Y_i)$.

PROOF We compute that

$$\begin{aligned} P(X > t) &= 1 - F_X(t) = 1 - \sum_i p_i F_i(t) \\ &= \sum_i p_i (1 - F_i(t)) = \sum_i p_i P(Y_i > t). \end{aligned}$$

Similarly,

$$P(X < t) = F_X(t^-) = \sum_i p_i F_i(t^-) = \sum_i p_i P(Y_i < t).$$

Hence, from Definition 3.7.1,

$$\begin{aligned} E(X) &= \int_0^\infty P(X > t)\, dt - \int_{-\infty}^0 P(X < t)\, dt \\ &= \int_0^\infty \sum_i p_i P(Y_i > t)\, dt - \int_{-\infty}^0 \sum_i p_i P(Y_i < t)\, dt \\ &= \sum_i p_i \left(\int_0^\infty P(Y_i > t)\, dt - \int_{-\infty}^0 P(Y_i < t)\, dt \right) \\ &= \sum_i p_i E(Y_i), \end{aligned}$$

as claimed. ∎

Summary of Section 3.7

- For general random variables, we can define a general expected value by $E(X) = \int_0^\infty P(X > t)\, dt - \int_{-\infty}^0 P(X < t)\, dt$.
- This definition agrees with our previous one, for discrete or absolutely continuous random variables.
- General expectation is still linear and monotone.

EXERCISES

3.7.1 Let X_1, X_2, and Y be as in Example 2.5.6, so that Y is a mixture of X_1 and X_2. Compute $E(X_1)$, $E(X_2)$, and $E(Y)$.

3.7.2 Suppose we roll a fair six-sided die. If it comes up 1, then we roll the same die again and let X be the value showing. If it comes up anything other than 1, then we

instead roll a fair eight-sided die (with the sides numbered 1 through 8), and let X be the value showing on the eight-sided die. Compute the expected value of X.

3.7.3 Let X be a positive constant random variable, so that $X = C$ for some constant $C > 0$. Prove directly from Definition 3.7.1 that $E(X) = C$.

3.7.4 Let Z be a general random variable (perhaps neither discrete nor continuous), and suppose that $P(Z \leq 100) = 1$. Prove directly from Definition 3.7.1 that $E(Z) \leq 100$.

3.7.5 Suppose we are told only that $P(X > x) = 1/x^2$ for $x \geq 1$, and $P(X > x) = 1$ for $x < 1$, but we are not told if X is discrete or continuous or neither. Compute $E(X)$.

3.7.6 Suppose $P(Z > z) = 1$ for $z \leq 5$, $P(Z > z) = (8 - z)/3$ for $5 < z < 8$, and $P(Z > z) = 0$ for $z \geq 8$. Compute $E(Z)$.

3.7.7 Suppose $P(W > w) = e^{-5w}$ for $w \geq 0$ and $P(W > w) = 1$ for $w < 0$. Compute $E(W)$.

3.7.8 Suppose $P(Y > y) = e^{-y^2/2}$ for $y \geq 0$ and $P(Y > y) = 1$ for $y < 0$. Compute $E(Y)$. (Hint: The density of a standard normal might help you solve the integral.)

3.7.9 Suppose the cdf of W is given by $F_W(w) = 0$ for $w < 10$, $F_W(w) = w - 10$ for $10 \leq w \leq 11$, and by $F_W(w) = 1$ for $w > 11$. Compute $E(W)$. (Hint: Remember that $F_W(w) = P(W \leq w) = 1 - P(W > w)$.)

CHALLENGES

3.7.10 Prove part (c) of Theorem 3.7.2. (Hint: If $X \leq Y$, then how does the event $\{X > t\}$ compare to the event $\{Y > t\}$? Hence, how does $P(X > t)$ compare to $P(Y > t)$? And what about $\{X < t\}$ and $\{Y < t\}$?)

3.8 Further Proofs (Advanced)

Proof of Theorem 3.4.7

We want to prove that if S has a compound distribution as in (3.4.2), then (a) $E(S) = E(X_1)\,E(N)$ and (b) $m_S(s) = r_N(m_{X_1}(s))$.

Because the $\{X_i\}$ are i.i.d., we have $E(X_i) = E(X_1)$ for all i. Define I_i by $I_i = I_{\{1,\ldots,N\}}(i)$. Then we can write $S = \sum_{i=1}^{\infty} X_i I_i$. Also note that $\sum_{i=1}^{\infty} I_i = N$.

Because N is independent of X_i, so is I_i, and we have

$$\begin{aligned} E(S) &= E\left(\sum_{i=1}^{\infty} X_i I_i\right) = \sum_{i=1}^{\infty} E(X_i I_i) \\ &= \sum_{i=1}^{\infty} E(X_i)E(I_i) = \sum_{i=1}^{\infty} E(X_1)E(I_i) \\ &= E(X_1)\sum_{i=1}^{\infty} E(I_i) = E(X_1)E\left(\sum_{i=1}^{\infty} I_i\right) \\ &= E(X_1)E(N). \end{aligned}$$

This proves part (a).

Now, using an expectation version of the law of total probability (see Theorem 3.5.3), and recalling that $E(\exp\left(\sum_{i=1}^{n} s\, X_i\right)) = m_{X_1}(s)^n$ because the $\{X_i\}$ are i.i.d., we compute that

$$\begin{aligned}
&m_S(s) \\
&= E\left(\exp\left(\sum_{i=1}^{n} sX_i\right)\right) = \sum_{n=0}^{\infty} P(N=n) E\left(\exp\left(\sum_{i=1}^{n} s\, X_i\right) \mid N=n\right) \\
&= \sum_{n=0}^{\infty} P(N=n) E\left(\exp\left(\sum_{i=1}^{n} s\, X_i\right)\right) = \sum_{n=0}^{\infty} P(N=n) m_{X_1}(s)^n \\
&= E(m_{X_1}(s)^N) = r_N(m_{X_1}(s)),
\end{aligned}$$

thus proving part (b). ∎

Proof of Theorem 3.5.3

We want to show that when X and Y are random variables, and $g : R^1 \to R^1$ is any function, then $E\,(g(Y)\,E(X \mid Y)) = E\,(g(Y)\,X)$.

If X and Y are discrete, then

$$\begin{aligned}
E\,(g(Y)\,E(X \mid Y)) &= \sum_{y \in R^1} g(y) E(X \mid Y=y)\, P(Y=y) \\
&= \sum_{y \in R^1} g(y) \left(\sum_{x \in R^1} x\, P(X=x \mid Y=y)\right) P(Y=y) \\
&= \sum_{y \in R^1} g(y) \left(\sum_{x \in R^1} x\, \frac{P(X=x,\ Y=y)}{P(Y=y)}\right) P(Y=y) \\
&= \sum_{x \in R^1} \sum_{y \in R^1} g(y) x\, P(X=x,\ Y=y) = E(g(Y)X),
\end{aligned}$$

as claimed.

Similarly, if X and Y are jointly absolutely continuous, then

$$\begin{aligned}
E\,(g(Y)\,E(X \mid Y)) &= \int_{-\infty}^{\infty} g(y)\, E(X \mid Y=y)\, f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} g(y) \left(\int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx\right) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} g(y) \left(\int_{-\infty}^{\infty} x \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx\right) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y)\, x\, f_{X,Y}(x,y)\, dx\, dy = E(g(Y)\,X),
\end{aligned}$$

as claimed. ∎

Proof of Theorem 3.5.4

We want to prove that, when X and Y are random variables, and $g : R^1 \to R^1$ is any function, then $E(g(Y)\,X \mid Y) = g(Y)\,E(X \mid Y)$.

For simplicity, we assume X and Y are discrete; the jointly absolutely continuous case is similar. Then for any y with $P(Y = y) > 0$,

$$\begin{aligned} E(g(Y)\,X \mid Y &= y) = \sum_{x \in R^1} \sum_{z \in R^1} g(z)\,x\,P(X = x, Y = z \mid Y = y) \\ &= \sum_{x \in R^1} g(y)\,x\,P(X = x \mid Y = y) \\ &= g(y) \sum_{x \in R^1} x\,P(X = x \mid Y = y) = g(y)\,E\,(X \mid Y = y)\,. \end{aligned}$$

Because this is true for any y, we must have $E(g(Y)X \mid Y) = g(Y)\,E(X \mid Y)$, as claimed. ∎

Proof of Theorem 3.5.6

We need to show that for random variables X and Y, $Var(X) = Var(E(X \mid Y)) + E\,(\mathrm{Var}(X \mid Y))$.

Using Theorem 3.5.2, we have that

$$\mathrm{Var}(X) = E((X - \mu_X)^2) = E(E((X - \mu_X)^2 \mid Y)). \tag{3.8.1}$$

Now,

$$\begin{aligned} (X - \mu_X)^2 &= (X - E(X \mid Y) + E(X \mid Y) - \mu_X)^2 \\ &= (X - E(X \mid Y))^2 + (E(X \mid Y) - \mu_X)^2 \\ &\quad + 2\,(X - E(X \mid Y))\,(E(X \mid Y) - \mu_X)\,. \end{aligned} \tag{3.8.2}$$

But $E\left((X - E(X \mid Y))^2 \mid Y\right) = \mathrm{Var}(X \mid Y)$.

Also, again using Theorem 3.5.2,

$$E(E((E(X \mid Y) - \mu_X)^2 \mid Y)) = E((E(X \mid Y) - \mu_X)^2) = \mathrm{Var}\,(E(X \mid Y))\,.$$

Finally, using Theorem 3.5.4 and linearity (Theorem 3.5.1), we see that

$$\begin{aligned} &E\,((X - E(X \mid Y))\,(E(X \mid Y) - \mu_X) \mid Y) \\ &= (E(X \mid Y) - \mu_X)\,E\,(X - E(X \mid Y) \mid Y) \\ &= (E(X \mid Y) - \mu_X)\,(E(X \mid Y) - E\,(E(X \mid Y) \mid Y)) \\ &= (E(X \mid Y) - \mu_X)\,(E(X \mid Y) - E(X \mid Y)) = 0. \end{aligned}$$

From (3.8.1), (3.8.2), and linearity, we have that $\mathrm{Var}(X) = E\,(\mathrm{Var}(X \mid Y))$ $+ \mathrm{Var}(E(X \mid Y)) + 0$, which completes the proof. ∎

Proof of Theorem 3.6.3 (*Cauchy–Schwartz inequality*)

We will prove that whenever X and Y are arbitrary random variables, each having finite, nonzero variance, then

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)\,\text{Var}(Y)}.$$

Furthermore, if Var(Y) > 0, then equality is attained if and only if $X - \mu_X = \lambda(Y - \mu_Y)$ *where* $\lambda = Cov(X, Y)/\,Var(Y)$.

If $\text{Var}(Y) = 0$, then Challenge 3.3.30 implies that $Y = \mu_Y$ with probability 1 (because $\text{Var}(Y) = E((Y - \mu_Y)^2) \geq 0$). This implies that

$$\text{Cov}\,(X, Y) = E\,((X - \mu_X)\,(\mu_Y - \mu_Y)) = 0 = \sqrt{\text{Var}\,(X)\ \text{Var}\,(Y)},$$

and the Cauchy–Schwartz inequality holds.

If $\text{Var}(Y) \neq 0$, let $Z = X - \mu_X$ and $W = Y - \mu_Y$. Then for any real number λ, we compute, using linearity, that

$$\begin{aligned} E\left((Z - \lambda W)^2\right) &= E(Z^2) - 2\lambda E(ZW) + \lambda^2 E(W^2) \\ &= \text{Var}(X) - 2\lambda\,\text{Cov}(X, Y) + \lambda^2\,\text{Var}(Y) \\ &= a\lambda^2 + b\lambda + c, \end{aligned}$$

where $a = \text{Var}(Y) > 0, b = -2\,\text{Cov}(X, Y)$, and $c = \text{Var}(X)$. On the other hand, clearly $E\left((Z - \lambda W)^2\right) \geq 0$ for all λ. Hence, we have a quadratic equation that is always nonnegative, and so has at most one real root.

By the quadratic formula, any quadratic equation has *two* real roots provided that the discriminant $b^2 - 4ac > 0$. Because that is not the case here, we must have $b^2 - 4ac \leq 0$, i.e.,

$$4\,\text{Cov}(X, Y)^2 - 4\,\text{Var}(Y)\,\text{Var}(X) \leq 0.$$

Dividing by 4, rearranging, and taking square roots, we see that

$$|\text{Cov}(X, Y)| \leq (\text{Var}(X)\,\text{Var}(Y))^{1/2},$$

as claimed.

Finally, $|\text{Cov}(X, Y)| = (\text{Var}(X)\,\text{Var}(Y))^{1/2}$ if and only if $b^2 - 4ac = 0$, which means the quadratic has one real root. Thus, there is some real number λ such that $E((Z - \lambda W)^2) = 0$. Since $(Z - \lambda W)^2 \geq 0$, it follows from Challenge 3.3.29 that this happens if and only if $Z - \lambda W = 0$ with probability 1, as claimed. When this is the case, then

$$\text{Cov}(X, Y) = E\,(ZW) = E(\lambda W^2) = \lambda E(W^2) = \lambda\,\text{Var}\,(Y)$$

and so $\lambda = \text{Cov}(X, Y)/\,\text{Var}(Y)$ when $\text{Var}(Y) \neq 0$. ∎