Definition 4.4:

Let X and Y be random variables with joint probability distribution $f(x, y)$. The covariance of X and Y is

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_y) f(x, y)$$

if X and Y are discrete, and

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_y) f(x, y)\, dx\, dy$$

if X and Y are continuous.

The covariance between two random variables is a measure of the nature of the association between the two. If large values of X often result in large values of Y or small values of X result in small values of Y, positive $X - \mu_X$ will often result in positive $Y - \mu_Y$ and negative $X - \mu_X$ will often result in negative $Y - \mu_Y$. Thus, the product $(X - \mu_X)(Y - \mu_Y)$ will tend to be positive. On the other hand, if large X values often result in small Y values, the product $(X - \mu_X)(Y - \mu_Y)$ will tend to be negative. The *sign* of the covariance indicates whether the relationship between two dependent random variables is positive or negative. When X and Y are statistically independent, it can be shown that the covariance is zero (see Corollary 4.5). The converse, however, is not generally true. Two variables may have zero covariance and still not be statistically independent. Note that the covariance only describes the *linear* relationship between two random variables. Therefore, if a covariance between X and Y is zero, X and Y may have a nonlinear relationship, which means that they are not necessarily independent.

The alternative and preferred formula for σ_{XY} is stated by Theorem 4.4.

Theorem 4.4: The covariance of two random variables X and Y with means μ_X and μ_Y, respectively, is given by

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y.$$

***Proof*:** For the discrete case, we can write

$$\begin{aligned}\sigma_{XY} &= \sum_x \sum_y (x-\mu_X)(y-\mu_Y)f(x,y)\\ &= \sum_x \sum_y xyf(x,y) - \mu_X \sum_x \sum_y yf(x,y)\\ &\quad - \mu_Y \sum_x \sum_y xf(x,y) + \mu_X\mu_Y \sum_x \sum_y f(x,y).\end{aligned}$$

Since

$$\mu_X = \sum_x xf(x,y), \quad \mu_Y = \sum_y yf(x,y), \text{ and } \sum_x \sum_y f(x,y) = 1$$

for any joint discrete distribution, it follows that

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y.$$

For the continuous case, the proof is identical with summations replaced by integrals. ∎

Example 4.13: Example 3.14 on page 95 describes a situation involving the number of blue refills X and the number of red refills Y. Two refills for a ballpoint pen are selected at random from a certain box, and the following is the joint probability distribution:

		x			
	$f(x,y)$	0	1	2	$h(y)$
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
	$g(x)$	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Find the covariance of X and Y.

***Solution*:** From Example 4.6, we see that $E(XY) = 3/14$. Now

$$\mu_X = \sum_{x=0}^{2} xg(x) = (0)\left(\frac{5}{14}\right) + (1)\left(\frac{15}{28}\right) + (2)\left(\frac{3}{28}\right) = \frac{3}{4},$$

and

$$\mu_Y = \sum_{y=0}^{2} yh(y) = (0)\left(\frac{15}{28}\right) + (1)\left(\frac{3}{7}\right) + (2)\left(\frac{1}{28}\right) = \frac{1}{2}.$$

Therefore,

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = \frac{3}{14} - \left(\frac{3}{4}\right)\left(\frac{1}{2}\right) = -\frac{9}{56}.$$

Example 4.14: The fraction X of male runners and the fraction Y of female runners who compete in marathon races are described by the joint density function

$$f(x,y) = \begin{cases} 8xy, & 0 \le y \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the covariance of X and Y.

***Solution*:** We first compute the marginal density functions. They are

$$g(x) = \begin{cases} 4x^3, & 0 \le x \le 1, \\ 0, & \text{elsewhere,} \end{cases}$$

and

$$h(y) = \begin{cases} 4y(1-y^2), & 0 \le y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

From these marginal density functions, we compute

$$\mu_X = E(X) = \int_0^1 4x^4\,dx = \frac{4}{5} \text{ and } \mu_Y = \int_0^1 4y^2(1-y^2)\,dy = \frac{8}{15}.$$

From the joint density function given above, we have

$$E(XY) = \int_0^1 \int_y^1 8x^2y^2\,dx\,dy = \frac{4}{9}.$$

Then

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = \frac{4}{225}.$$

Although the covariance between two random variables does provide information regarding the nature of the relationship, the magnitude of σ_{XY} *does not indicate anything regarding the strength of the relationship*, since σ_{XY} is not scale-free. Its magnitude will depend on the units used to measure both X and Y. There is a scale-free version of the covariance called the **correlation coefficient** that is used widely in statistics.

Definition 4.5: Let X and Y be random variables with covariance σ_{XY} and standard deviations σ_X and σ_Y, respectively. The correlation coefficient of X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

It should be clear to the reader that ρ_{XY} is free of the units of X and Y. The correlation coefficient satisfies the inequality $-1 \le \rho_{XY} \le 1$. It assumes a value of zero when $\sigma_{XY} = 0$. Where there is an exact linear dependency, say $Y \equiv a + bX$,

$\rho_{XY} = 1$ if $b > 0$ and $\rho_{XY} = -1$ if $b < 0$. (See Exercise 4.48.) The correlation coefficient is the subject of more discussion in Chapter 12, where we deal with linear regression.

Example 4.15: Find the correlation coefficient between X and Y in Example 4.13.
***Solution*:** Since

$$E(X^2) = (0^2)\left(\frac{5}{14}\right) + (1^2)\left(\frac{15}{28}\right) + (2^2)\left(\frac{3}{28}\right) = \frac{27}{28}$$

and

$$E(Y^2) = (0^2)\left(\frac{15}{28}\right) + (1^2)\left(\frac{3}{7}\right) + (2^2)\left(\frac{1}{28}\right) = \frac{4}{7},$$

we obtain

$$\sigma_X^2 = \frac{27}{28} - \left(\frac{3}{4}\right)^2 = \frac{45}{112} \text{ and } \sigma_Y^2 = \frac{4}{7} - \left(\frac{1}{2}\right)^2 = \frac{9}{28}.$$

Therefore, the correlation coefficient between X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-9/56}{\sqrt{(45/112)(9/28)}} = -\frac{1}{\sqrt{5}}.$$

Example 4.16: Find the correlation coefficient of X and Y in Example 4.14.
***Solution*:** Because

$$E(X^2) = \int_0^1 4x^5\, dx = \frac{2}{3} \text{ and } E(Y^2) = \int_0^1 4y^3(1-y^2)\, dy = 1 - \frac{2}{3} = \frac{1}{3},$$

we conclude that

$$\sigma_X^2 = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75} \text{ and } \sigma_Y^2 = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = \frac{11}{225}.$$

Hence,

$$\rho_{XY} = \frac{4/225}{\sqrt{(2/75)(11/225)}} = \frac{4}{\sqrt{66}}.$$

Note that although the covariance in Example 4.15 is larger in magnitude (disregarding the sign) than that in Example 4.16, the relationship of the magnitudes of the correlation coefficients in these two examples is just the reverse. This is evidence that we cannot look at the magnitude of the covariance to decide on how strong the relationship is.

Exercises

4.33 Use Definition 4.3 on page 120 to find the variance of the random variable X of Exercise 4.7 on page 117.

4.34 Let X be a random variable with the following probability distribution:

x	-2	3	5
$f(x)$	0.3	0.2	0.5

Find the standard deviation of X.

4.35 The random variable X, representing the number of errors per 100 lines of software code, has the following probability distribution:

x	2	3	4	5	6
$f(x)$	0.01	0.25	0.4	0.3	0.04

Using Theorem 4.2 on page 121, find the variance of X.

4.36 Suppose that the probabilities are 0.4, 0.3, 0.2, and 0.1, respectively, that 0, 1, 2, or 3 power failures will strike a certain subdivision in any given year. Find the mean and variance of the random variable X representing the number of power failures striking this subdivision.

4.37 A dealer's profit, in units of \$5000, on a new automobile is a random variable X having the density function given in Exercise 4.12 on page 117. Find the variance of X.

4.38 The proportion of people who respond to a certain mail-order solicitation is a random variable X having the density function given in Exercise 4.14 on page 117. Find the variance of X.

4.39 The total number of hours, in units of 100 hours, that a family runs a vacuum cleaner over a period of one year is a random variable X having the density function given in Exercise 4.13 on page 117. Find the variance of X.

4.40 Referring to Exercise 4.14 on page 117, find $\sigma^2_{g(X)}$ for the function $g(X) = 3X^2 + 4$.

4.41 Find the standard deviation of the random variable $g(X) = (2X + 1)^2$ in Exercise 4.17 on page 118.

4.42 Using the results of Exercise 4.21 on page 118, find the variance of $g(X) = X^2$, where X is a random variable having the density function given in Exercise 4.12 on page 117.

4.43 The length of time, in minutes, for an airplane to obtain clearance for takeoff at a certain airport is a random variable $Y = 3X - 2$, where X has the density function

$$f(x) = \begin{cases} \frac{1}{4}e^{-x/4}, & x > 0 \\ 0, & \text{elsewhere.} \end{cases}$$

Find the mean and variance of the random variable Y.

4.44 Find the covariance of the random variables X and Y of Exercise 3.39 on page 105.

4.45 Find the covariance of the random variables X and Y of Exercise 3.49 on page 106.

4.46 Find the covariance of the random variables X and Y of Exercise 3.44 on page 105.

4.47 For the random variables X and Y whose joint density function is given in Exercise 3.40 on page 105, find the covariance.

4.48 Given a random variable X, with standard deviation σ_X, and a random variable $Y = a + bX$, show that if $b < 0$, the correlation coefficient $\rho_{XY} = -1$, and if $b > 0$, $\rho_{XY} = 1$.

4.49 Consider the situation in Exercise 4.32 on page 119. The distribution of the number of imperfections per 10 meters of synthetic failure is given by

x	0	1	2	3	4
$f(x)$	0.41	0.37	0.16	0.05	0.01

Find the variance and standard deviation of the number of imperfections.

4.50 For a laboratory assignment, if the equipment is working, the density function of the observed outcome X is

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

Find the variance and standard deviation of X.

4.51 For the random variables X and Y in Exercise 3.39 on page 105, determine the correlation coefficient between X and Y.

4.52 Random variables X and Y follow a joint distribution

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

Determine the correlation coefficient between X and Y.