4.3. The Correlation Coefficient

1. *Definition.* Suppose that X_1 and X_2 are random variables. Denote

$$\mu_1 = E(X_1), \qquad \mu_2 = E(X_2),$$

and denote

$$\sigma_1^2 = Var(X_1), \qquad \sigma_2^2 = Var(X_2).$$

The covariance of X_1 and X_2 is

$$\begin{aligned} \sigma_{12} = Cov(X_1, X_2) &= E((X_1 - \mu_1)(X_2 - \mu_2)) \\ &= E(X_1 X_2) - E(X_1)E(X_2). \end{aligned}$$

If $\sigma_1 > 0$ and $\sigma_2 > 0$, then the correlation of X_1 and X_2 is

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

Here, σ_1 is the standard deviation of X_1, σ_2 is the standard deviation of X_2.

2. *A Special case.* If X is a random variable, then

$$Cov(X,X) = Var(X).$$

3. And, it is clear that

$$Cov(X,Y) = Cov(Y,X).$$

4. *Example.* Suppose that X and Y are discrete random variables, and their joint pmf is given by the table below.

$X\backslash f\backslash Y$	1	2
1	0.4	0.1
2	0.1	0.4

(a) Find the correlation of X and Y.

(b) Find $Cov(X, Y)$.

— *Hint.* Just a reminder that the table is a summary of the following information:

$$P(X = 1, Y = 1) = 0.4, \quad P(X = 1, Y = 2) = 0.1,$$

$$P(X = 2, Y = 1) = 0.1, \quad P(X = 2, Y = 2) = 0.4.$$

— *Solution.* First, let us find $E(XY)$. By definition, $E(XY)$ is a sum:

$$\begin{aligned} E(XY) &= 1 \times 1 \times 0.4 + 1 \times 2 \times 0.1 \\ &\quad + 2 \times 1 \times 0.1 + 2 \times 2 \times 0.4. \end{aligned}$$

In this sum, each term is a possible value of X times a possible value of Y times the corresponding probability. Calculation shows that

$$E(XY) = 2.4.$$

Denote by $f_2(y)$ the marginal pmf of Y. Then

y	1	2
$f_2(y)$	0.5	0.5

Hence, $E(Y) = 1.5$, $Var(Y) = 0.25$.

In a similar way, we can show that $E(X) = 1.5$, $Var(X) = 0.25$.

Finally, by the definition, we have

$$Cov(X, Y) = E(XY) - E(X)E(Y) = 2.4 - 1.5^2 = 0.15,$$

and, the correlation of X and Y is

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{0.15}{\sqrt{0.25}\ \sqrt{0.25}} = 0.6.$$

So, there is a positive (and somewhat strong) correlation between X and Y.

5. *Example.* Suppose X and Y are continuous random variables, and their joint pdf is

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

Here, following our convention, the trivial parts of the joint distribution are not given in this definition (1).

(a) Find $Cov(X, Y)$.

(b) Find the correlation of X and Y.

— *Solution.* First, let us find $E(XY)$. By definition, $E(XY)$ is a double integral:

$$E(XY) = \int_0^1 \int_0^1 xy(x+y)dydx$$

In this double integral, the integrand is x times y times the corresponding bit of probability, which is

$$f(x,y)dxdy = (x+y)dxdy.$$

Calculation shows that

$$E(XY) = \int_0^1 \frac{x\ (3x+2)}{6} dx = \tfrac{1}{3}.$$

We have shown that the marginal pdf of X is

$$f_1(x) = x + \tfrac{1}{2}, \quad 0 < x < 1.$$

Calculation shows that

$$E(X) = \frac{7}{12}, \quad E(X^2) = \frac{5}{12},$$

and

$$Var(X) = \frac{11}{144}.$$

Similarly,

$$E(Y) = \frac{7}{12}, \quad Var(Y) = \frac{11}{144}.$$

Finally, by the definition, we have

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{3} - \frac{7}{12} \cdot \frac{7}{12} = -\frac{1}{144},$$

and, the correlation of X and Y is

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{-\frac{1}{144}}{\sqrt{\frac{11}{144}}\sqrt{\frac{11}{144}}} = -\frac{1}{11} \approx -0.0909.$$

So, there is a negative (and somewhat weak) correlation between X and Y.

6. We will show later in the text that, if X and Y are independent, then the correlation between X and Y is $\rho = 0$.