2.5.3. Covariance and Variance of Sums of Random Variables

The covariance of any two random variables X and Y, denoted by $\text{Cov}(X, Y)$, is defined by

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

Note that if X and Y are independent, then by Proposition 2.3 it follows that $\text{Cov}(X, Y) = 0$.

Let us consider now the special case where X and Y are indicator variables for whether or not the events A and B occur. That is, for events A and B, define

$$X = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise,} \end{cases} \qquad Y = \begin{cases} 1, & \text{if } B \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$$

Then,

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

and, because XY will equal 1 or 0 depending on whether or not both X and Y equal 1, we see that

$$\text{Cov}(X, Y) = P\{X = 1, Y = 1\} - P\{X = 1\}P\{Y = 1\}$$

From this we see that

$$\begin{aligned}
\text{Cov}(X, Y) > 0 &\Leftrightarrow P\{X = 1, Y = 1\} > P\{X = 1\}P\{Y = 1\} \\
&\Leftrightarrow \frac{P\{X = 1, Y = 1\}}{P\{X = 1\}} > P\{Y = 1\} \\
&\Leftrightarrow P\{Y = 1|X = 1\} > P\{Y = 1\}
\end{aligned}$$

That is, the covariance of X and Y is positive if the outcome $X = 1$ makes it more likely that $Y = 1$ (which, as is easily seen by symmetry, also implies the reverse).

In general it can be shown that a positive value of $\text{Cov}(X, Y)$ is an indication that Y tends to increase as X does, whereas a negative value indicates that Y tends to decrease as X increases.

The following are important properties of covariance.

Properties of Covariance

For any random variables X, Y, Z and constant c,

1. $\text{Cov}(X, X) = \text{Var}(X)$,
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$,
3. $\text{Cov}(cX, Y) = c\,\text{Cov}(X, Y)$,
4. $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$.

Whereas the first three properties are immediate, the final one is easily proven as follows:

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

The fourth property listed easily generalizes to give the following result:

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j) \tag{2.15}$$

A useful expression for the variance of the sum of random variables can be obtained from Equation (2.15) as follows:

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j<i} \text{Cov}(X_i, X_j) \qquad (2.16)\end{aligned}$$

If $X_i, i = 1, \ldots, n$ are independent random variables, then Equation (2.16) reduces to

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

Definition 2.1 If $X_1, \dots, X_n$ are independent and identically distributed, then the random variable $\bar{X} = \sum_{i=1}^{n} X_i/n$ is called the *sample mean*.

The following proposition shows that the covariance between the sample mean and a deviation from that sample mean is zero. It will be needed in Section 2.6.1.

Proposition 2.4 Suppose that $X_1, \dots, X_n$ are independent and identically distributed with expected value μ and variance σ^2. Then,

(a) $E[\bar{X}] = \mu$.
(b) $\mathrm{Var}(\bar{X}) = \sigma^2/n$.
(c) $\mathrm{Cov}(\bar{X}, X_i - \bar{X}) = 0,\ i = 1, \dots, n$.

Proof Parts (a) and (b) are easily established as follows:

$$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{m} E[X_i] = \mu,$$

$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

To establish part (c) we reason as follows:

$$\begin{aligned}
\mathrm{Cov}(\bar{X}, X_i - \bar{X}) &= \mathrm{Cov}(\bar{X}, X_i) - \mathrm{Cov}(\bar{X}, \bar{X}) \\
&= \frac{1}{n}\mathrm{Cov}\left(X_i + \sum_{j \neq i} X_j, X_i\right) - \mathrm{Var}(\bar{X}) \\
&= \frac{1}{n}\mathrm{Cov}(X_i, X_i) + \frac{1}{n}\mathrm{Cov}\left(\sum_{j \neq i} X_j, X_i\right) - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0
\end{aligned}$$

where the final equality used the fact that X_i and $\sum_{j \neq i} X_j$ are independent and thus have covariance 0. ■

Equation (2.16) is often useful when computing variances.

Example 2.33 (Variance of a Binomial Random Variable) Compute the variance of a binomial random variable X with parameters n and p.