XXIX. Covariance, Correlation, and Linear Functions

Definition and Formulas for Covariance

- **Definition**: The covariance of two random variables Y_1 and Y_2 is

$$\text{Cov}(Y_1, Y_2) := E\left[(Y_1 - \mu_1)(Y_2 - \mu_2)\right],$$

  where μ_1 and μ_2 are the means of Y_1 and Y_2, respectively.

- Useful formulas to calculate covariance:

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= E(Y_1 Y_2) - E(Y_1) E(Y_2) \\ \text{Cov}(Y_1, Y_1) &= V(Y_1) \\ \text{Cov}(cY_1, cY_2) &= c^2 \,\text{Cov}(Y_1, Y_2) \end{aligned}$$

Intuition for Covariance

- Covariance is a measure of dependence.
- Dependence doesn't necessarily mean that the variables do the same thing; it means that knowing the value of one gives you more information about the other.
- If Y_1 moves with Y_2, then Cov is positive.
- If Y_1 moves consistently against Y_2, then Cov is negative.

Either one indicates dependence! It's like a child who always does the opposite of what you say: You can still control him by reverse psychology, so he is still dependent on you.

Independence Theorem

- **Theorem**: If Y_1 and Y_2 are independent, then $\text{Cov}(Y_1, Y_2) = 0$.
- The converse is not true:
- It is possible to have $\text{Cov}(Y_1, Y_2) = 0$ with Y_1 and Y_2 dependent.

See Example II for a counterexample.

Correlation Coefficient

- Let Y_1 and Y_2 be random variables with standard deviations σ_1 and σ_2.
- Define the correlation coefficient ρ:

$$\rho := \frac{\text{Cov}(Y_1, Y_2)}{\sigma_1 \sigma_2}$$

- Then $\rho(cY_1, cY_2) = \rho(Y_1, Y_2)$. So ρ is scale-independent.
- We always have $-1 \leq \rho \leq 1$.

Linear Functions of Random Variables

- Suppose $Y_1, \ldots, Y_n$ are random variables with means $\mu_1, \ldots, \mu_n$, variances $\sigma_1^2, \ldots, \sigma_n^2$.

1. The expected value $E\left(a_1Y_1 + \cdots + a_nY_n\right)$ is

$$a_1\mu_1 + \cdots + a_n\mu_n.$$

2. The variance $V\left(a_1Y_1 + \cdots + a_nY_n\right)$ is

$$a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 + 2\sum_{i>j} a_ia_j \operatorname{Cov}\left(Y_i, Y_j\right).$$

Linear Functions of Random Variables

- Suppose $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_m$ are random variables.

3. The covariance of $a_1Y_1 + \cdots + a_nY_n$ and $b_1X_1 + \cdots + b_mX_m$ is

$$\operatorname{Cov}\left(\sum a_iY_i, \sum b_jX_j\right) = \sum_{i,j} a_ib_j \operatorname{Cov}\left(Y_i, X_j\right).$$

Example I

Let $f\left(y_1, y_2\right) :\equiv 1$ over the triangle with corners at $(-1, 0)$,$(0, 1)$, and $(1, 0)$. Calculate $E\left(Y_1\right), E\left(Y_2\right)$ and $E\left(Y_1Y_2\right)$.

(Graph.)

$$\begin{aligned}
\mu_1 &:= E(Y_1) \\
&:= \int_{y_2=0}^{y_2=1} \int_{y_1=y_2-1}^{y_1=1-y_2} y_1 \cdot 1 \, dy_1 \, dy_2 \\
&= \cdots \\
&= 0 \quad \text{(We should have known this from symmetry.)} \\
\mu_2 &:= E(Y_2) \\
&:= \int_{y_2=0}^{y_2=1} \int_{y_1=y_2-1}^{y_1=1-y_2} y_2 \cdot 1 \, dy_1 \, dy_2 \\
&= \cdots \\
&= \frac{1}{3} \quad \text{This seems reasonable. Engineers might recognize it as the \underline{moment} of a triangle.} \\
E(Y_1 Y_2) &:= \int_{y_2=0}^{y_2=1} \int_{y_1=y_2-1}^{y_1=1-y_2} y_1 y_2 \cdot 1 \, dy_1 \, dy_2 \\
&= \cdots \\
&= 0 \quad \text{Still not surprising from symmetry.}
\end{aligned}$$

Example II

As in Example I, let $f(y_1, y_2) :\equiv 1$ over the triangle with corners at $(-1, 0)$,$(0, 1)$,$(1, 0)$. Compute $\text{Cov}(Y_1, Y_2)$. Are Y_1 and Y_2 independent?

(Graph.) $\text{Cov}(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1) E(Y_2) = 0 \cdot \frac{1}{3} - 0 = \boxed{0}$.
But we know Y_1 and Y_2 are not independent, because the region is not rectangular. (For example, if you know $Y_1 = \frac{3}{4}$, then you know Y_2 cannot be $\frac{1}{2}$. [Graph.]) So our theorem above is not iff.

Example III

Let Y_1 and Y_2 be independent variables with means and variances $\mu_1 = 7$, $\mu_2 = 5$, $\sigma_1^2 = 4$, $\sigma_2^2 = 9$. Let $U_1 := Y_1 + 2Y_2$ and $U_2 := Y_1 - Y_2$. Calculate $V(U_1)$ and $V(U_2)$.

$V(U_1) = 4 + 2^2 9 + 2 \cdot 1 \cdot 2 \,\text{Cov}(Y_1, Y_1) = 40 + 0 = \boxed{40}$ by independence. $V(U_2) = 4 + (-1)^2 9 + 2 \cdot 1 \cdot 2 \,\text{Cov}(Y_1, Y_1) = 13 + 0 = \boxed{13}$ by independence.

Example IV

As in Example III, let Y_1 and Y_2 be independent variables with means and variances $\mu_1 = 7$, $\mu_2 = 5$, $\sigma_1^2 = 4$, $\sigma_2^2 = 9$. Let $U_1 := Y_1 + 2Y_2$ and $U_2 := Y_1 - Y_2$. Calculate $\text{Cov}(U_1, U_2)$ and $\rho(U_1, U_2)$.

(a)

$$\begin{aligned}
\operatorname{Cov}(U_1, U_2) &= \operatorname{Cov}(Y_1, Y_1) - \operatorname{Cov}(Y_1, Y_2) + 2\operatorname{Cov}(Y_2, Y_1) - 2\operatorname{Cov}(Y_2, Y_2) \\
&= V(Y_1) - 0 + 2 \cdot 0 - 2V(Y_2) \\
&= 4 - 2 \cdot 9 \\
&= \boxed{-14}
\end{aligned}$$

(b) $\rho(U_1, U_2) \quad = \quad \dfrac{-14}{\sqrt{40}\sqrt{13}} \quad =$

$$\boxed{-\frac{7}{\sqrt{130}} \approx -0.614}$$

Note that $-1 \leq \rho \leq 1$.

Example V

Suppose Y_i are independent variables with mean μ and variance σ^2. Find the mean and variance of the average $\overline{Y} := \frac{1}{n}Y_1 + \cdots + \frac{1}{n}Y_n$.

(a) $E\left(\overline{Y}\right) = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \boxed{\mu}$. No surprise.

(b)

$$\begin{aligned} V\left(\overline{Y}\right) &= a_1^2 V\left(Y_1\right) + \cdots + a_n^2 V\left(Y_n\right) + 2\sum_{i>j} a_i a_j \operatorname{Cov}\left(Y_i, Y_j\right) \\ &= \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 + 2\sum_{i>j}\left(\frac{1}{n}\right)\left(\frac{1}{n}\right)\operatorname{Cov}\left(Y_i, Y_j\right) \end{aligned}$$

Because they are $\underline{\text{independent}}$, then $\operatorname{Cov}(Y_i, Y_j) = 0$.

$$= \boxed{\frac{\sigma^2}{n}}$$