LECTURE - 21

Agenda:

1. General definition of independence
2. Covariance and correlation
3. Examples

GENERAL DEFINITION OF INDEPENDENCE

Definition: If X and Y are two random variables, then X and Y are said to be independent if for every function g of X, and h of Y,

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

In the last lectures, we defined independence for a pair of discrete random variables, and also for a pair of continuous random variables. The above definition generalizes the notion of independence for an arbitrary pair of random variables.

COVARIANCE AND CORRELATION

Definition: The covariance between two random variables X and Y is defined as

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))].$$

Just as $E(X)$, $V(X)$ help us understand the probabilist behaviour of a single random variable X in specific ways, $Cov(X,Y)$ helps us understand the joint probabilistic behaviour of two random variables X and Y as follows:

(1) If Y tends to be large when X is large, and Y tends to be small when X is small, then X and Y have a positive covariance

(2) If Y tends to be large when X is small, and Y tends to be small when X is large, then X and Y have a negative covariance.

Here are two properties of $Cov(X,Y)$ which are useful for practical purposes.

(i) $Cov(X,Y) = E[XY] - E(X)\,E(Y)$

(ii) $Cov(X,X) = V(X)$

<u>Definition:</u> The correlation coefficient between two random variables X and Y is defined as

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)\,V(Y)}}.$$

Properties of ρ_{XY}

(1) ρ_{XY} is a unitless quantity

(2) $-1 \leq \rho_{XY} \leq 1$

(3) If $\rho_{XY} = -1$, $Y = aX + b$ where $a < 0$.

(4) If $\rho_{XY} = +1$, $Y = aX + b$ where $a > 0$.

(5) ρ_{XY} is a measure of "linear association" between X and Y. As ρ_{XY} becomes closer to zero, it indicates lesser and lesser linear relationship between X and Y.

<u>Result:</u> If X and Y are independent,

$$\rho_{XY} = Cov(X, Y) = 0.$$

<u>Proof:</u> Choose $g(x) = x$ and $h(y) = y$. Then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)] \quad \text{(by independence)}$$

Hence,

$$E[XY] - E(X)E(Y) = 0.$$

However, the converse is not true, i.e.,

$$\text{cov}(X, Y) = 0 \not\Longrightarrow X \text{ and } Y \text{ are independent}$$

Example: Consider discrete random variables X and Y with joint p.m.f. given by

$$P(X = x, Y = y) = \begin{cases} \frac{1}{8} & \text{if } (x, y) \neq (0,0), \\ 0 & \text{if } (x, y) = 0 \\ \mathcal{X} = \{-1, 0, 1\}, & \\ \mathcal{Y} = \{-1, 0, 1\}. & \end{cases}$$

$$E[XY] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (xy)\, P(X = x, Y = y)$$

$$= 0$$

$$E[X] = \sum_{x \in \mathcal{X}} x\, P(X = x) = 0$$

$$E(Y) = \sum_{y \in \mathcal{Y}} y\, P(Y = y) = 0$$

Hence, $E(XY) - E(X)E(Y) = 0$.

However,

$$P(X=-1, Y=-1) = \frac{1}{8}$$

$$P(X=-1)P(Y=-1) = \frac{3}{8}\frac{3}{8} = \frac{9}{64}$$

Hence,

$$P(X=-1, Y=-1) \neq P(X=-1)P(Y=-1),$$

which means X and Y are not independent.

Suppose we have a random experiment which gives rise to two random variables X and Y. Suppose we are interested in $V(X-Y)$. Here is a general identity which helps us compute this quantity in terms of $V(X)$, $V(Y)$ and $Cov(X,Y)$.

<u>Result:</u> Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two sequences of random variables. Let us define

$$U_1 = \sum_{i=1}^{n} a_i X_i \quad , \quad U_2 = \sum_{j=1}^{m} b_j Y_j .$$

Then,

(1) $$V(U_1) = \sum_{i=1}^{n} a_i^2 V(X_i) + \sum\sum_{1 \le i \ne j \le n} a_i a_j Cov(X_i, X_j)$$

(2) $$Cov(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j).$$

<u>Example:</u> Suppose that random variables X and Y have the joint density

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2} & \text{if } 0 \le x \le y, \quad 0 \le y \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Calculate $V(Y-X)$.

$$\boxed{V(Y-X) = V(Y) + V(X) + 2(1)(-1)\,Cov(X,Y)}$$

(Use the previous result with $U_1 = 1.Y + (-1)X$, i.e., $a_1 = 1$, $a_2 = -1$, $X_1 = Y$, $X_2 = X$, $n = 2$)

$$E(X) = \int_0^2 \int_0^y x \cdot \frac{1}{2}\, dx\, dy = \int_0^2 \frac{y^2}{4} dy = \frac{2}{3}.$$

$$E(Y) = \int_0^2 \int_0^y y \cdot \frac{1}{2}\, dx\, dy = \int_0^2 \frac{y^2}{2} dy = \frac{4}{3}.$$

$$E(X^2) = \int_0^2 \int_0^y x^2 \cdot \frac{1}{2}\, dx\, dy = \frac{2}{3}.$$

$$E(Y^2) = \int_0^2 \int_0^y y^2 \cdot \frac{1}{2}\, dx\, dy = 2.$$

$$E(XY) = \int_0^2 \int_0^y xy\, dx\, dy = \int_0^2 \frac{y^3}{4} dy = 1.$$

Using these values,

$$V(X) = E(X^2) - (E(X))^2 = \frac{2}{9}.$$

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{2}{9}$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{9}.$$

Hence,

$$V(Y-X) = \frac{2}{9} + \frac{2}{9} - 2\times\frac{1}{9} = \frac{2}{9}.$$

We have been using informally that if $X_1, X_2, \ldots, X_n$ are random variables which arise independently, then $V(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} V(X_i)$. Now, we are in a position to see a formal proof of this fact. Note that,

$$\begin{aligned} V\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} V(X_i) + \sum\sum_{1\le i\ne j\le n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} V(X_i) + \sum\sum_{1\le i\ne j\le n} 0 \\ &= \sum_{i=1}^{n} V(X_i). \end{aligned}$$