Basic Statistics 1

Class 1

Wonmun Shin

(wonmun.shin@sejong.ac.kr)

Department of Economics, Sejong University

* This lecture note is written based on Professor Chang Sik Kim's lecture note .

Random Variables

Random Variables

- **Random variable (r.v.)**: a variable whose numerical value is determined by the outcome of a random experiment
  - Discrete random variable: can take on *finite* or *countably infinite* number of values
  - Continuous random variable: can take *any* value in an interval
- **Probability Density Function (pdf)**
  - Suppose that X is a discrete r.v. and the x is one of its possible values.
  - Denote $P(X = x)$ as the probability that the r.v. takes the specific value x.
  - Then, the discrete r.v. has a probability density function (pdf) which represents the probabilities for all the possible outcomes.

$$f(x) = \begin{cases} P(X = x_i) & \text{for } i = 1, 2, 3, \cdots, n, \cdots \\ 0 & \text{for } X \neq x_i \end{cases}$$

Random Variables [cont'd]

- **Properties**
  1. $P(X = x_i) \geq 0$ for $i = 1, 2, \cdots, n, \cdots$
  2. $\sum_x f(x) = 1$

- Then, what if X is a continuous variable?

- $f(x)$ is a pdf of a continuous r.v. X if

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

$$\int_a^b f(x)\, dx = P(a \leq x \leq b)$$

- Note $\quad P(X = a) = \int_a^a f(x) dx = 0$

Mean and Variance

Expected Values (Expectation)

- The probability distribution contains all the information about the r.v., but in some cases it is desirable to have some numerical summary measures of the distribution's characteristics such as *mean* and *variance*.
- Definition of $E(X)$
  - Expected value of a r.v. X = Mean of X = First moment of X
  - Measure of central location of X
  - Average of values that X can take, weighted by corresponding probabilities
- Discrete r.v. case

$$\mu = E(X) = \sum_{x} x f(x)$$

- Continuous r.v. case

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$$

Expected Values (Expectation) [cont'd]

- **Useful results**
  - $E(a) = a$
    - <u>Why?</u> $E(a) = \sum_x af(x) = a\underbrace{\sum_x f(x)}_{=1} = a$
  - $E(aX) = aE(X)$
    - <u>Why?</u> $E(aX) = \sum_x (ax) f(x) = a\underbrace{\sum_x xf(x)}_{=E(X)} = aE(X)$
  - $E(aX + b) = aE(X) + b$
    - <u>Why?</u> $E(aX + b) = \sum_x (ax + b) f(x) = a\underbrace{\sum_x xf(x)}_{=E(X)} + b\underbrace{\sum_x f(x)}_{=1} = aE(X) + b$

Variance (and Standard Deviation)

- We can consider the measure of dispersion in the pdf of a r.v.. This is a weighed average of the squared discrepancy about the mean.
- Definition of $Var\,(X)$
  - Variance of a r.v. $X =$ Second moment of X
  - Measure of dispersion of X around $E\,(X)$

$$\sigma^2 = Var\,(X) = E\,(X - \mu)^2$$

- Standard deviation $= \sigma = \sqrt{Var\,(X)}$

Variance (and Standard Deviation) [cont'd]

- **Alternative** (and useful) **expression**

$$Var\left(X\right) = E\left(X^2\right) - \mu^2$$

  - Why?

$$\begin{aligned} E\left(X-\mu\right)^2 &= E\left(X^2 - 2\mu X + \mu^2\right) = \sum_x \left(x^2 - 2\mu x + \mu^2\right) f\left(x\right) \\ &= \sum_x x^2 f\left(x\right) + \sum_x \left(-2\mu x\right) f\left(x\right) + \sum_x \mu^2 f\left(x\right) \\ &= E\left(X^2\right) - 2\mu E\left(X\right) + \mu^2 \\ &= E\left(X^2\right) - \mu^2 \end{aligned}$$

Variance (and Standard Deviation) [cont'd]

- **Useful results**
  - $Var\left(a\right) = 0$
    - Why?

$$Var\left(a\right) = E\left(a^2\right) - \left[E\left(a\right)\right]^2 = a^2 - a^2 = 0$$

  - $Var\left(X + a\right) = Var\left(X\right)$
    - Why?

$$\begin{aligned} Var\left(X + a\right) &= E\left[\left(X + a\right)^2\right] - \left[E\left(X + a\right)\right]^2 \\ &= E\left(X^2 + 2aX + a^2\right) - \left[\mu^2 + 2a\mu + a^2\right] \\ &= E\left(X^2\right) + 2aE\left(X\right) + a^2 - \mu^2 - 2a\mu - a^2 \\ &= E\left(X^2\right) + 2a\mu + a^2 - \mu^2 - 2a\mu - a^2 \\ &= E\left(X^2\right) - \mu^2 = Var\left(X\right) \end{aligned}$$

- **Useful results [Cont'd]**
  - $Var\,(aX + b) = a^2 Var\,(X)$
    - Why?

$$
\begin{aligned}
Var\,(aX+b) &= Var\,(aX) \\
&= E\left(a^2X^2\right) - \left[E\,(aX)\right]^2 \\
&= a^2E\left(X^2\right) - (a\mu)^2 \\
&= a^2\left[E\left(X^2\right) - \mu^2\right] \\
&= a^2\,Var\,(X)
\end{aligned}
$$

Covariance

Two Random Variables

- Suppose that X and Y are random variables.

$$E(X \pm Y) = E(X) \pm E(Y)$$

$$E(aX \pm bY) = aE(X) \pm bE(Y)$$

$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2 \cdot Cov(X, Y)$$

$$Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2ab \cdot Cov(X, Y)$$

Covariance

- What is $Cov\,(X, Y)$?
  - **Covariance** measures the linear relationship between two random variables. For example, we can check whether higher values of X are associated with higher values of Y or not.

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

  - Positive σ_{XY}: High values of X tend to be associated with high values of Y.
  - Negative σ_{XY}: High values of X tend to be associated with low values of Y.
  - $\sigma_{XY} = 0$: No linear relationship (possible non-linear relationship)
- Note Existence of covariance does **NOT** imply any **causality** between two.

Correlation Coefficient

- **Correlation coefficient**
  - Since the magnitude of covariance does not tell us the degree of strength of linear dependence, we need to *normalize* the value of the covariance by dividing variances of X and Y.

$$\rho_{XY} = \frac{Cov\left(X, Y\right)}{\sqrt{Var\left(X\right)} \cdot \sqrt{Var\left(Y\right)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho_{XY} \leq 1$$

Independence

- If X and Y are independent,

$$E(XY) = E(X)\,E(Y)$$

  - Why?

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy \underbrace{f_{xy}(x,y)}_{joint\ pdf} \\ &= \sum_x \sum_y xyf(x)\,f(y) \\ &= \sum_x xf(x) \sum_y yf(y) \\ &= E(X)\,E(Y) \end{aligned}$$

- Therefore, when X and Y are independent,

$$Cov(X,Y) = 0$$

- **Note** The converse is not necessarily true. That is, $Cov(X,Y) = 0$ does **not always** imply X and Y are independent.