Module 18

CHARACTERISTICS OF A PROBABILITY DISTRIBUTION

- X: a r.v. with support S_X, d.f. $F_X(\cdot)$ and p.m.f./p.d.f. $f_X(\cdot)$;

- The probability distribution (i.e. d.f./p.m.f./p.d.f.) of X describes the manner in which the r.v. X takes different values in $\mathbb{R}$;

- It may be desirable to have a set of numerical measures that provide a summary of prominent features of the probability distribution of X.

A. Measures of Central Tendency or Measures of Location

A measure of central tendency gives us the idea about the central value of the distribution around which values of r.v. X are clustered. Some of the measures of central tendency are;

(a) Mean:

$$\mu = E(X) = \begin{cases} \sum_{x \in S_X} x \ f_X(x), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x \ f_X(x) dx, & \text{if } X \text{ is A.C.} \end{cases}.$$

(b) Median: Median m of the distribution is a value such that the r.v. X is equally likely to be either smaller or larger than m. One may define median as

$$m = \inf\{x \in \mathbb{R} : F_X(x) \geq \frac{1}{2}\}.$$

Clearly, for continuous (in particular for A.C.) r.v.

$$F_X(m) = P(\{X \leq m\}) = \frac{1}{2}.$$

It can be shown that, for any r.v. X,

$$E(|X - m|) \leq E(|X - c|), \quad \forall\ c \in \mathbb{R}.$$

(c) Mode: Roughly Speaking the mode of distribution of X is the value that is most likely to occur. One may define mode m_0 of distribution of X as

$$m_0 = \arg \sup_{x \in \mathbb{R}} f_X(x).$$

A probability distribution may have more than one modes (bimodal, trimodal and multimodal distributions).

(d) Geometric Mean (G.M.) If $P(X > 0) = 1$, then the G.M. G of X is defined by

$$\ln G = E(\ln X),$$

i.e.,

$$G = e^{E(\ln X)}.$$

(e) Harmonic Mean (H.M.).The H.M of X is defined by

$$\frac{1}{H} = E\left(\frac{1}{X}\right),$$

i.e.,

$$H = \frac{1}{E\left(\frac{1}{X}\right)}.$$

- Using the Jensen inequality it can be shown that

$$A.M. \geq G.M. \geq H.M.$$

- For unimodal symmetric distributions, generally one has

$$\text{mean} \;=\; \text{median} \;=\; \text{mode.}$$

(B) Measures of Dispersion (or Variation)

Measures of central tendency give us the idea about the location of central part of the distribution. Other measures are often needed to describe probability distribution. A measure of dispersion (or variation) gives us an idea of the scatter (or cluster or dispersion) of the values of random variable X about a measure of central tendency. Various measures of dispersion are:

(a) Mean Deviation

Let A be a suitable measure of central tendency. The mean deviation (M.D.) of X about A is defined by

$$MD(A) = E(|X - A|).$$

Generally, $A = \mu = E(X)$, in which case $MD(\mu)$ is called the mean deviation about mean. It can be shown that

$$MD(m) \leq MD(A), \quad \forall\ A \in \mathbb{R},$$

where m is the median.

(b) Standard Deviation and Variance

The standard deviation (s.d.) of X is defined by

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{E((X-\mu)^2)},$$

where $\mu = E(X)$. Note that the s.d. is in the same units as the r.v. X and it measures the average distance of values of X from the mean μ. For any $A \in \mathbb{R}$

$$\begin{aligned} E((X-A)^2) &= E((X-\mu+\mu-A)^2) \\ &= E((X-\mu)^2) + (\mu-A)^2 \\ &\geq E((X-\mu)^2) = \mathrm{Var}(X). \end{aligned}$$

(c) Quartile Deviation

- For a fixed $p \in (0,1)$, the quantity

$$\xi_p = \inf\{x \in \mathbb{R} : F_x(x) \geq p\}$$

is called the quantile of order p.

- We call

$$\begin{aligned} Q_1 &= \xi_{\frac{1}{4}} = \text{first quartile;} \\ Q_2 &= \xi_{\frac{1}{2}} = \text{second quartile (or median)} \\ Q_3 &= \xi_{\frac{3}{4}} = \text{third quartile} \end{aligned}$$

Clearly Q_1,Q_2 and Q_3 divide the probability mass of distribution into four equal parts each containing 25% weight.

- The quantity

$$Q = \frac{Q_3 - Q_1}{2}$$

  is called the quartile deviation (or semi interquartile range) and it gives us the idea about the scatter of the probability mass of the distribution.

(d) Coefficient of Variation

$$\mathrm{Var}(X) = E((X-\mu)^2), \quad \text{where } \mu = E(X),$$

is not independent of units and is therefore not appropriate for comparing variability of two r.v.s. having different measurement units. In such situations one may use the coefficient of variation, defined as,

$$C_X = \frac{\sqrt{Var(X)}}{E(X)} = \frac{\sigma}{\mu} \qquad \text{(variation per unit mean)}$$

(C) Measures of Skewness

- For unimodal distributions, a measure of skewness gives us the idea about departure from symmetry;
- Positively skewed p.m.f/p.d.f. has more probability mass to the right of mode with longer right tail;
- Negatively skewed p.m.f./p.d.f. has more probability mass to the left of mode with longer left tail;
- For symmetric distributions (since $X - \mu \overset{d}{=} \mu - X$)

$$\mu_3 = E((X - \mu)^3) = 0.$$

(a) The quantity

$$\beta_1 = \frac{E((X-\mu)^3)}{[E((X-\mu)^2)]^{\frac{3}{2}}} \text{ (independent of Units)}$$

is taken as a measure of skewness and is called coefficient of skewness. For symmetrical distribution, $\beta_1 = 0$.

(b) For symmetric distribution ($X - \mu \overset{d}{=} \mu - X$) it can be shown that

$$Q_3 - m = m - Q_1$$

$$\Leftrightarrow Q_3 - 2m + Q_1 = 0$$

Thus the quantity

$$\beta_2 = \frac{Q_3 - 2m + Q_1}{Q_3 - Q_1}$$

may also be taken as a measure of skewness, and is called the Yule coefficient of skewness. The quantity $Q_3 - Q_1$ in the denominator of β_2 is to make it independent of units. For positively (negatively) skewed distribution we have $\beta_2 > \ (<) \ 0$.

(D) Measures of Kurtosis

(a) The quantity

$$\gamma_1 = \frac{\mu_4}{\mu_2^2} = \frac{E((X-\mu)^4)}{[E((X-\mu)^2)]^2}$$

is called the kurtosis of probability distribution of X.

(b) The kurtosis of an unimodal distribution gives us the idea about flatness or peakedness of the distribution around mode.

(c) For a standard unimodal symmetric distribution (called the normal distribution) the kurtosis is 3.

(d) The quantity

$$\gamma_2 = \gamma_1 - 3$$

is called the excess Kurtosis of distribution of X.

(e) Mesokurtic distributions: $\gamma_2 = 0$;

- Leptokurtic distribution: $\gamma_2 > 0$ and distribution has sharper peak and longer flatter tails;

- Platykurtic distribution: $\gamma_2 < 0$ and distribution has flatter peak and shorter thinner tails.

Take Home Problem

- Let $\{X_\theta : \theta > 0\}$ be a family r.v.s with X_θ having p.d.f.

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}.$$

Compare different features of probability distributions of r.v.s X_{θ_1} and X_{θ_2}, where $0 < \theta_1 < \theta_2$.

Abstract of Next Module

- In many situations we may be interested in studying two or more numerical characteristics of the sample space simultaneously. This leads to studying joint probability distribution of two or more random variables defined on the same sample space.

- In the next module we will study joint distribution of two or more random variables.

Thank you for your patience