

Instructors: Purushottam Kar, Neeraj Misra
Authors: Gurpreet Singh
Email: guggu@iitk.ac.in

CHARACTERISTICS OF A PROBABILITY DISTRIBUTION

1. Measures of Central Tendency

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 given below.

1. Mean / Average

$$\mu = \mathbb{E}[X] = \begin{cases} \sum_{x \in S_X} x f_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{if } X \text{ is a.c.} \end{cases}$$

2. 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

$$m = \inf \left\{ x \in \mathbb{R} \mid F_X(x) \geq \frac{1}{2} \right\}$$

For continuous (in particular a.c.) r.v.

$$F_X(m) = \mathbb{P}[X \leq m] = \frac{1}{2}$$

Also, it can be shown that for any r.v. X

$$\mathbb{E}[|X - m|] \leq \mathbb{E}[|X - c|]$$

3. Mode 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

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

Remark. For unimodal symmetric distributions, mean = median = mode.

4. Geometric Mean If $X > 0$ almost surely (a.s.), then the G.M. G of X is defined by

$$\ln(G) = \mathbb{E}[\ln(X)]$$

5. Harmonic Mean The H.M. H of X is defined by

$$\frac{1}{H} = \mathbb{E}\left[\frac{1}{X}\right]$$

Remark. Using the Jensen Inequality, we can say $\mu > G > H$

2. Measures of Dispersion / 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 as follows.

- 1. Mean Deviation** For some $A \in \mathbb{R}$, such that A is a suitable measure of central tendency (generally μ), the mean deviation (MD) of X about A is defined by

$$\text{MD}(A) = \mathbb{E}[|X - A|]$$

- 2. Standard Deviation and Variance** The standard deviation (s.d.) of X is defined by

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\mathbb{E}[(X - \mu)^2]}$$

where $\mu = \mathbb{E}[X]$. 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 $A \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}[(X - A)^2] &= \mathbb{E}[(X - \mu + \mu - A)^2] \\ &= \mathbb{E}[(X - \mu)^2] + (\mu - A)^2 + 2(\mu - A) \cdot \mathbb{E}[X - \mu] \\ &\geq \mathbb{E}[(X - \mu)^2] = \text{Var}(X)\end{aligned}$$

- 3. Quartile Deviation** For a fixed $p \in (0, 1)$, the quantity

$$\xi_p = \inf \{x \in \mathbb{R} \mid F_X(x) \geq p\}$$

is called the quantile of order p . We define

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

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

- 4. Coefficient of Variation** The variance term 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 follows.

$$C_X = \frac{\sqrt{\text{Var}(X)}}{\mathbb{E}[X]} = \frac{\sigma}{\mu}$$

3. Measures of Skewness

For unimodal distributions, a measure of skewness gives us the idea about skewness from symmetry. If the probability distribution is

- (i) positively skewed, then the p.m.f / p.d.f has more probability mass to the right of mode with longer right tail
- (ii) negatively skewed, then the p.m.f / p.d.f has more probability mass to the left of mode with longer left tail

Various measures of skewness are given below.

1. Coefficient of Skewness The quantity

$$\beta_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$

where μ_r is the r^{th} moment. For symmetric distributions, $\beta_1 = 0$

2. Yule Coefficient of Skewness For symmetric distributions ($X - \mu = \mu - X$), it can be shown that

$$\begin{aligned} Q_3 - m &= m - Q_1 \\ Q_3 - 2m + Q_1 &= 0 \end{aligned}$$

Thus the quantity

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

is also taken as a measure of skewness.

4. Measure of Kurtosis

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

$$\gamma_1 = \frac{\mu_4}{\mu_2^2}$$

is called the kurtosis of probability distribution of X .

For a standard unimodal symmetric distribution (called the normal distribution, discussed later) the kurtosis is 3. Hence $\gamma_2 = \gamma_1 - 3$ is known as the excess Kurtosis of distribution of X . For distributions where

- (i) $\gamma_2 = 0$, the distribution is called a Mesokurtic Distribution
- (ii) $\gamma_2 > 0$, the distribution is called a Leptokurtic Distribution, and the distribution has sharper peak and longer flatter tails
- (iii) $\gamma_2 < 0$, the distribution is called a Platykurtic Distribution, and the distribution has flatter peak and shorter thinner tails