16-July-2025
Wednesday

⊙ EDA

1. Measure of Central Tendency

a. Mean
b. Median
c. Mode

Central Tendency: The property of data being concentrated in the centre

a 1.[illegible] Km
b 1.3 Km
c. 1.5 Km
d. 1.1 Km
e. 15 Km

(here the 15 is outliers we exclude to avoid wrong mean)

∴ If there is no outliers then we use mean

Mean: It is the average of all numbers and is sometimes called as Arithmetic mean

Median: The statistical median is the middle number in a sequence of numbers.

Mode: The mode is the number that occur often with a set of number

- Discrete numerical data is go with mode.
- Used for categorical data mostly

- Mode Example:

$3\ 7\ \underline{5}\ 2\underline{5}$ Mode = 5 (Most Frequent)

⊙ Measure of Spread / Data Variability

- Range.
  - The difference btw the highest and lowest value within a set of numbers

- Interquartile range (IQR)
  - The interquartile data is the middle half of the data

Q_1 - 25 %

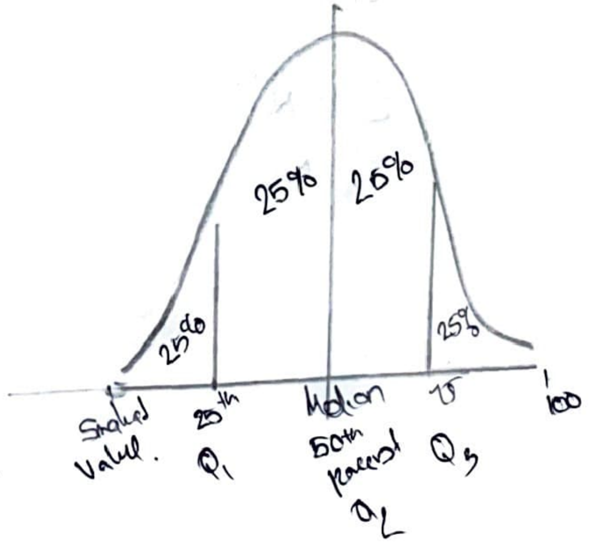
Q_2 - 50 %

Q_3 - 75 %

∴ Q_2 is the median

$$1\ \underbrace{2\ 3}_{\substack{2.5\\Q_1}}\ 4\ \underset{Q_2}{\boxed{5}}\ 6\ \underbrace{7\ 8}_{\substack{75\\Q_3}}\ 9$$

Figure:



$$IQR = Q_3 - Q_1$$
$$= 7.5 - 2.5$$
$$= \underline{\underline{5}}$$

∴ Q_2 is not always equal to IQR

another example:

$$1\ \underbrace{2\quad 3}_{2.5}\ 4\ \underbrace{5\quad 6}_{5.5}\ 7\ \underbrace{8\ \ 9}_{8.5}\ 10$$

$$IQR = 8.5 - 2.5$$
$$= \underline{\underline{6}}$$

- Standard Deviation (σ)

The standard deviation is the measure that is used to quantify the amount of variation or dispers of a set of data values

Population

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Sample

$$S = \sqrt{\frac{\Sigma(X-\bar{x})^2}{n-1}}$$

standard deviation (σ) = $\sqrt{\text{Variance}}$

· square root of variance is "SD"

- Variance

The variance is a measure of how far a set of data are dispersed out from their mean/average value

· The 'σ^2' is variance.

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \qquad S^2 = \frac{\Sigma(X-M)^2}{N-1}$$

Example

x	$x-\mu$ or $x-\bar{x}$	$(x-\mu)^2$
1	$1-3=-2$	4
2	$2-3=-1$	1
3	0	0
4	1	1
5	2	4

$$\text{Variance} = \frac{\Sigma(x-\mu)^2}{n} = \frac{10}{5} = 2$$

$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = \underline{\underline{3}}$$

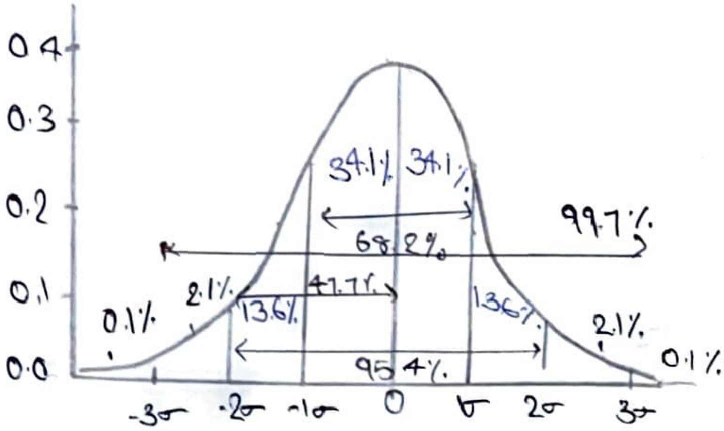
Variance $= 2$

and

$$SD = \sqrt{\text{Variance}} = \sqrt{2} = \underline{\underline{1.414}}$$

- Percentile / "Centile"

Fig.



A percentile is a measure used in statistics indicating the value below which a given percentage of observation in a group of observations fall.

⊙ Distributions

The graphical representation of all observations is also known as distribution.

Types of distribution

Continuous Distribution

Discrete Distribution

- Normal Distribution

Normal distribution, also known as the Gaussian dis

It is a probability distribution that is symmetric about the mean.

- Which follow 3 σ rule
- Most prefered in the pipeline
- Bell shaped Curve.

- Properties of Normal/Gaussian dist

1 Emperical rule
2 · Distortion is Normal dist
3 · Central Limit theorem.
4 · Standard Normal dist
5 · Outlies
6 · QQ plot
7 · Log, sqrt, Boxcox transformation

1. Empirical rule

- Figure:

<u>Same figures as the previous percentile</u>

. The empirical rule states that for a normal distribution nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

· 68 % data fall within the 1st SD from the mean

·95% fall within 2 SD

.99.7% fall within 3 SD

·Any point lying after 3 sigma is outlier.

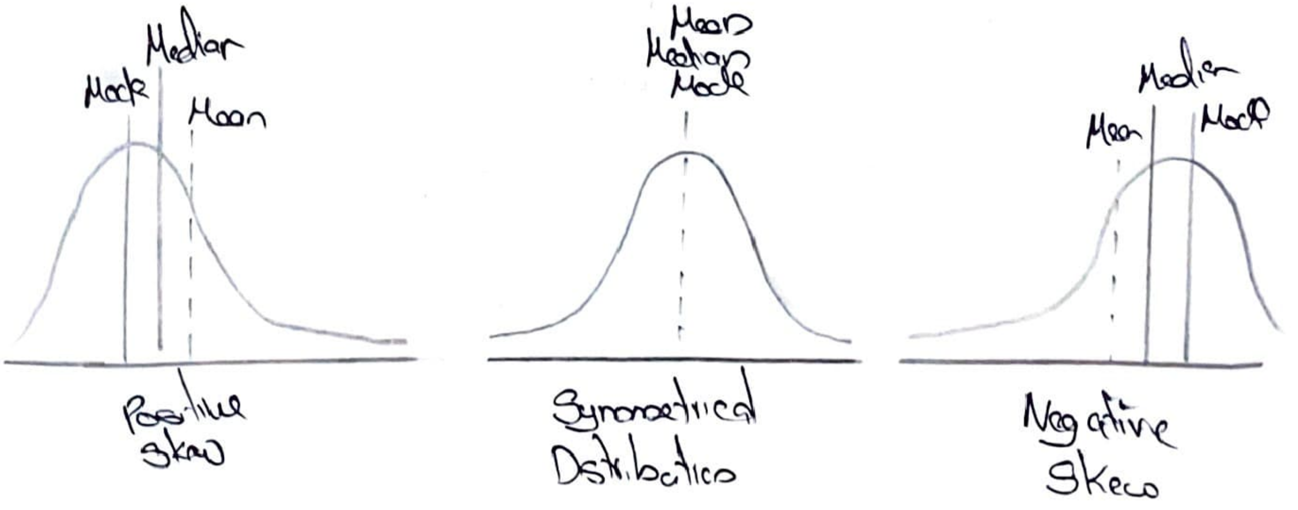
2. Distortion is Normal Distribution

The distortion is normally distributed curves can be quantified in 2 ways.

1. skewness
2. kurtosis

1. skewness in Normal Distribution

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution



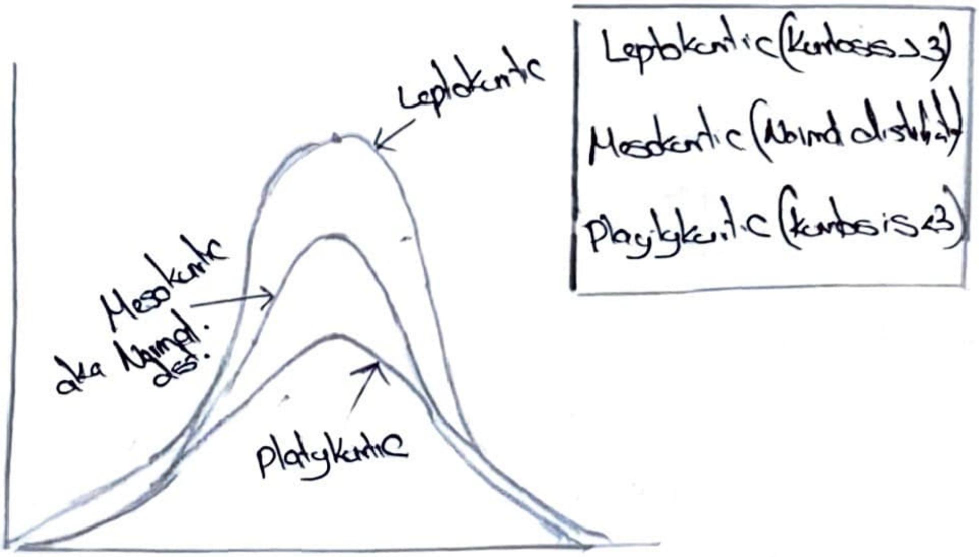
- How much skewness and kurtosis

  - If the skewness is between -0.5 and 0.5 → fairly Symme
  - If -1 and -0.5 or 0.5 and 1 → moderately skewed
  - If skewness is > 1 or < -1 → highly skewed

- A standard normal distribution has 3 kurtosis.

- Kurtosis.

In probability theory and statistics, kurtosis is a measure of the "peakedness" of the probability distribution of an real-valued random variable


Leptokurtic(kurtosis > 3)
Mesokurtic (Normal distribution)
Platykurtic (kurtosis < 3)
Leptokurtic
Mesokurtic
aka Normal dist.
Platykurtic

- Which is Best - the Mean, Median, or Mode?

  - When you have a symmetrical distribution for continuous data, the mean, median and mode are equal. In this case, analysts tend to use the mean because it include all of the data in the calculations. However, if you have a skewed distribution, the median is often the best measure of central tendency

- When you have categorical or discrete data, the median or mode is usually the best choice
For categorical data, you have to use the mode

Preferable:

Normal dist → Mean

Skewed dist → Median

Categorical dist → Mode

- Correction Distortion is Normal Distribution

Transformation is nothing but taking a mathematical function and applying it to the data.

Log Transformation [Each data point is replaced with log(x) to obtain ND]

Square-root Transformation [Each data point is replaced by its square root]

Reciprocal Transformation [It takes the inverse of x ie, $1/x$]

Box-Cox Transformation. [Transformation of non-normal dependent variable to normal shape]

Reason: To transform the data to either reduce the skewness or to normalize the data or simplify make the data easier to understand

- Outliers.

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from that data.

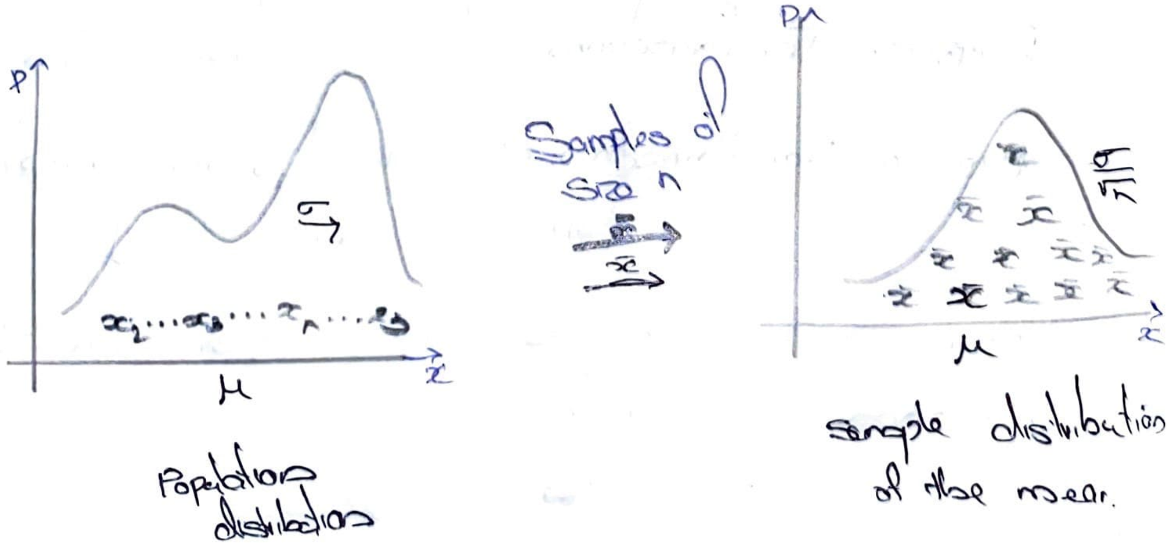
⊙ Central limit theorem

— The central theorem states that the distribution of sample means approximates a normal distribution as the sample size get larger, regardless of population distribution shape

- CLT in one sentence "Even if X_n not normal, the avg is normal"

- When ~~in one~~ collecting mean the samples from any dist the no of samples taken for calculating the mean should be greater or equal to 30

- Rule of thumb: the sample should be bigger than 25 observations or ("30")


P
$\bar{x}$
$x_1 \cdots x_2 \cdots x_n \cdots x_3$
μ
x
Samples of size n
$\bar{x}$
P
$\frac{\sigma}{\sqrt{n}}$
$\bar{x}$
μ
x

Population distribution

Sample distribution of the mean.