Statistics

A sample is a set of individuals selected from a population - we usually want samples to be representative.

2 rules - They have to be → Representative
↳ Generalized

Parameter & Statistic :

Parameter describes the population whereas statistic describes a sample.

Statistics → Descriptive
Statistics → Inferential

Constants : Charecterstic that is fixed across conditions
~~Variable~~ ex:- State, Country, Planets etc

Variable : A charecterstic that changes across conditions.
ex:- salary, temperature, respiratory rate etc.

Types of Variables:

Independent : One that is manipulated.

Dependent : One that is observed.

Relationship V/s Causal effects :

① Correlational Methods : They only tell how directly or inversely are variables related, but cannot demonstrate cause-and-effect relationship aka Causality.

② Experimental methods :- Here, the independent variables are manipulated, whereas the dependent variable is observed. eg :- Vaccine research on humans.

③ Quasi Experimental Method :- It aims to establish a cause-and-effect relationship between an independent and dependent variable.

Scales of Measurement :-

* By putting them into categories (qualitative)
* By using numbers (quantitative)

2 types of measurement's variables fall into :-

1. discrete
2. Continous

as well as different **measurement scales** :-

- nominal
- Ordinal

} Categorical (data are counted)

- Interval
- Ratio

} Quantitative (data that are measured)

Shape of a Frequency distributions :-

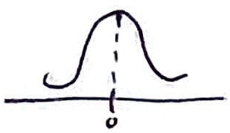
Disussed different graphs

↳ Bar ~~chart~~ Graph
↳ Pie chart
↳ Histogram
↳ Polygon etc.

N → Total no. of samples in the population

n → no. of sample's collected in a sample.

Normal / Symmetric distribution :-



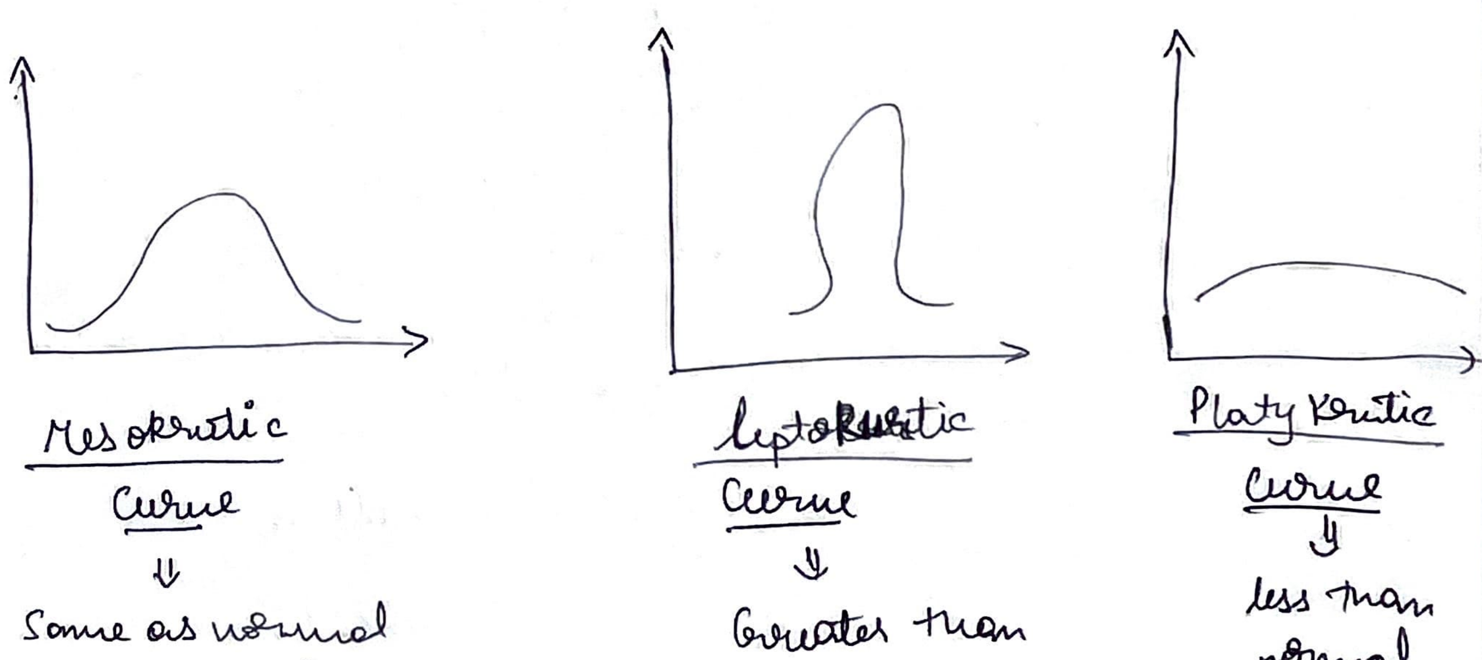
More the number of samples the closer the distribution looks symmetrical in nature

Nearly all distributions can be classified as either symmetrical or skewed.

Tail on the right side

Tail on the left.

Kurtosis :- The sharpness of a frequency distribution curve.



Mesokurtic Curve ⇓ Same as normal distribution

Leptokurtic Curve ⇓ Greater than normal distribution

Platykurtic Curve ⇓ less than normal dist.

Discussed Rank (OR) Percentile :- Percentage of individual under a specific value.

Percentile Rank -> Refers to percentage

Percentile -> Refers to a score.

Quantiles :-

Divide distributions into equal parts

Centiles :- Divide into 100 equal parts

Quartiles :- 4 equal parts

Quintiles :- 5 equal parts

Deciles :- 10 equal parts

Central Tendency :-

3 characteristics to describe any distributions :-

* Shape
* Central Tendency (where is the center located)
* Variability (how spread out the scores are?)

3 methods to determine C.T :- Mean, Median, Mode

1. Mean (μ) :

$$\boxed{\mu = \frac{\Sigma x}{N}}$$

Mean of Population

$$\boxed{\bar{x} = \frac{\Sigma x}{n}}$$

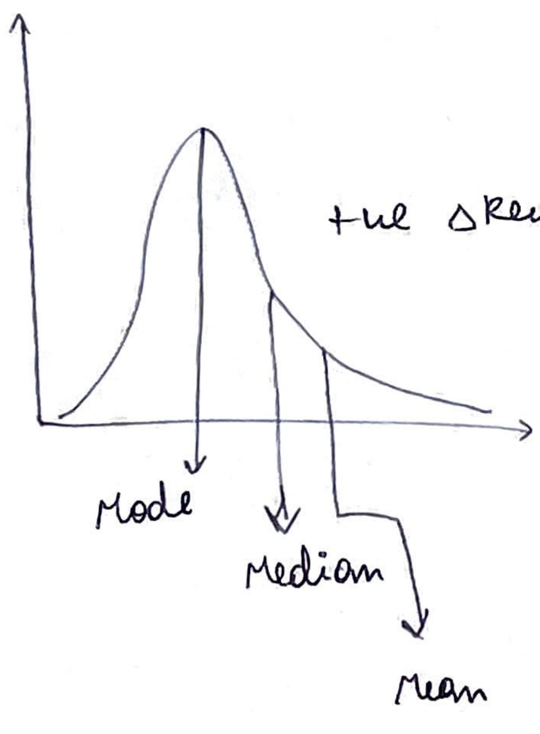
Mean of sample.

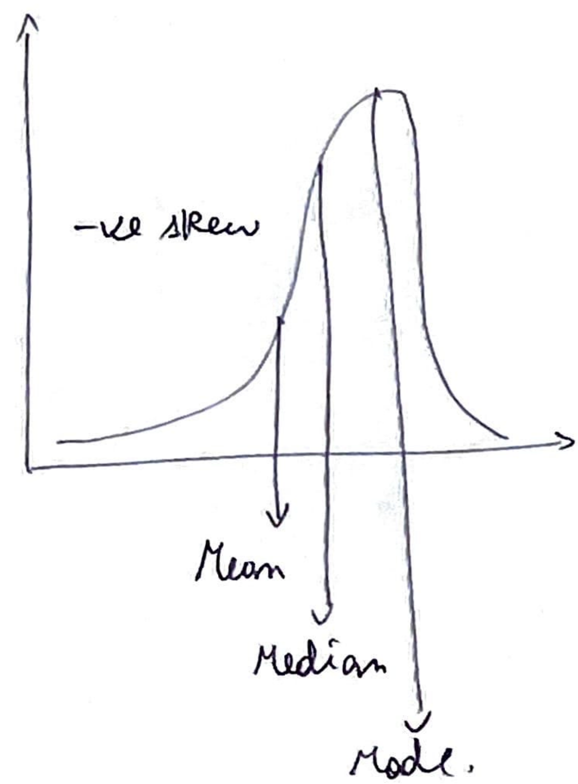
The median :

Divides the distribution exactly in half when all scores / data are arranged in ascending order.

The Mode : The most common score in the sample of data.

* when distribution have 2 modes (Bi-model) & 3 modes (tri-model).





Data Variability :

Method to measure and objectively describe the differences that exist within a dataset - in statistics that measure is called data variability.

1) Range

2) Interquartile Range

3) Standard Deviation. *

<u>Range</u> : The dist. b/w the largest score and smallest score in a distribution

$$\text{Range} = (\text{Highest } X - \text{lowest } X) + 1$$

* Sensetive to large / extreme scores.

<u>Interquartile Range:</u>

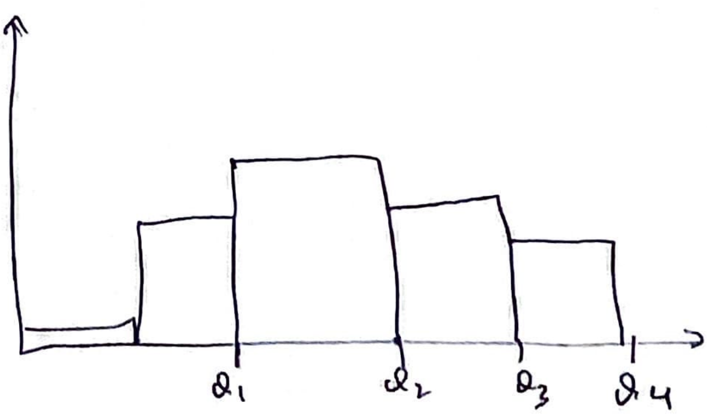
This avoids extreme scores, instead it measures the range covered by the middle 50%.

Q1 - 1st Quartile - 25th % (C_{25})

Q2 - 2nd " —— " - 50th % (C_{50})

Q3 - 3rd " —— " - 75th % (C_{75})

Q4 - 4th " —— " - 100 % (C_{100})



hence, Interquartile Range = $Q_3 - Q_1$

still disregards 50% of the scores, hence not providing the complete picture of variability.

Standard Deviation :-

The average distence of the scores from the mean.

* Uses mean as the reference point. It approximates the average distence from the mean.

Charecterstics of Standard deviation:

* measures "typical" (standard) distence from the mean
* It allows us to visualize the distributions
* smaller the S.D, the more accurately a sample will represent its population.

Deviation score = $X - \mu$

next we find the mean of the deviation scores,

$$\Sigma (X - \mu) = 0$$

because of the problem of the sum of deviation scores leading to zero due to the ⅀ -ve values.

Solution is squaring each deviation score, from which we calculate mean squared deviation. (Variance)

The final step is to make a correction for having squared all the distances

$$\boxed{\text{Standard Deviation} = \sqrt{\text{Variance}}}$$

SS = Sum of Squares

① $\boxed{SS = \Sigma (x - \mu)^2}$ (Derived formula)

② $\boxed{SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}}$ (computational formula)

Population Standard Deviation :-

$$\boxed{\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{\sum_{i=1}^{N} (X - \mu)^2}{N}}}$$

Population Variance :-

$$\Rightarrow \boxed{\sigma^2 = \frac{SS}{N} = \frac{\sum_{i=1}^{N} (x - \mu)^2}{N}}$$

Shows relationship b/w Standard Deviation and Variance.

To make sample variability unbiased, we change n to n-1.

$$\boxed{S = \sqrt{\frac{SS}{n-1}}.}$$

Sample Standard Deviation

$$\boxed{S^2 = \frac{SS}{n-1}}$$

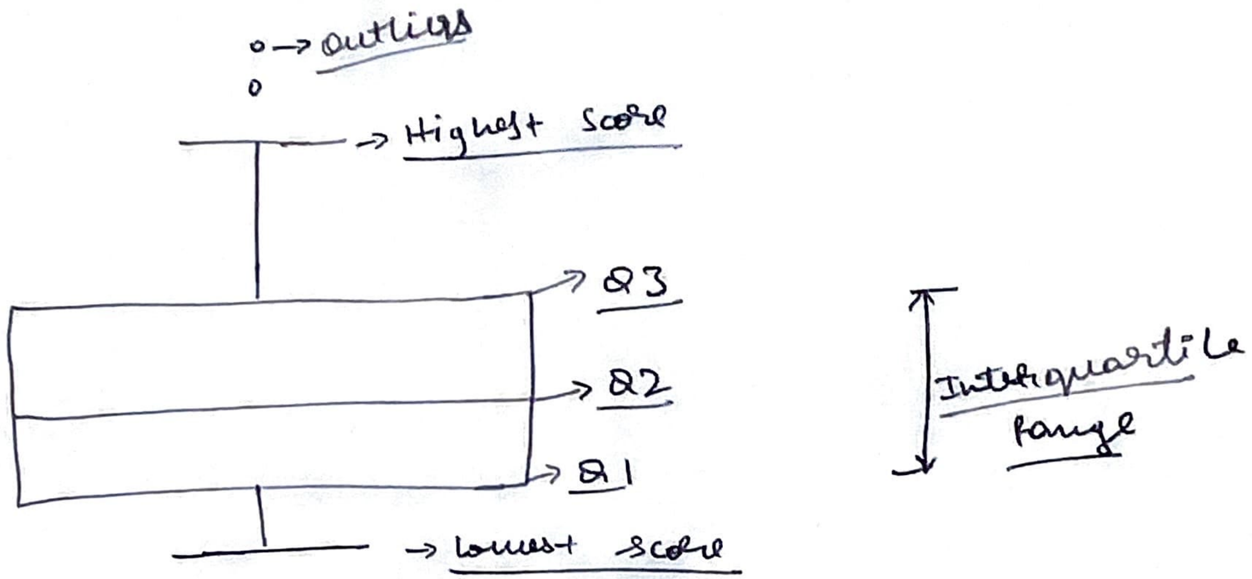
sample Variance

"n-1", corrects for the bias, compensating for the undersampling.

Visual display's of Variability :

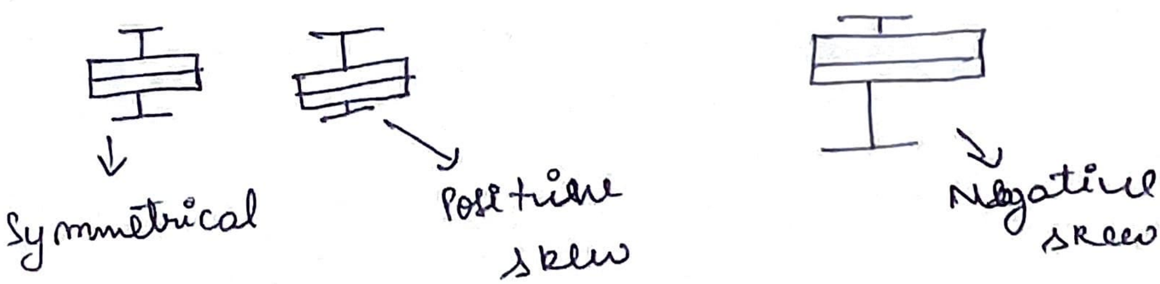
① Error Bar

② **Boxplots :-** This splits the data into quartiles.



Range = Highest - Lowest

Box plots also provide insights about the shape of the data set.

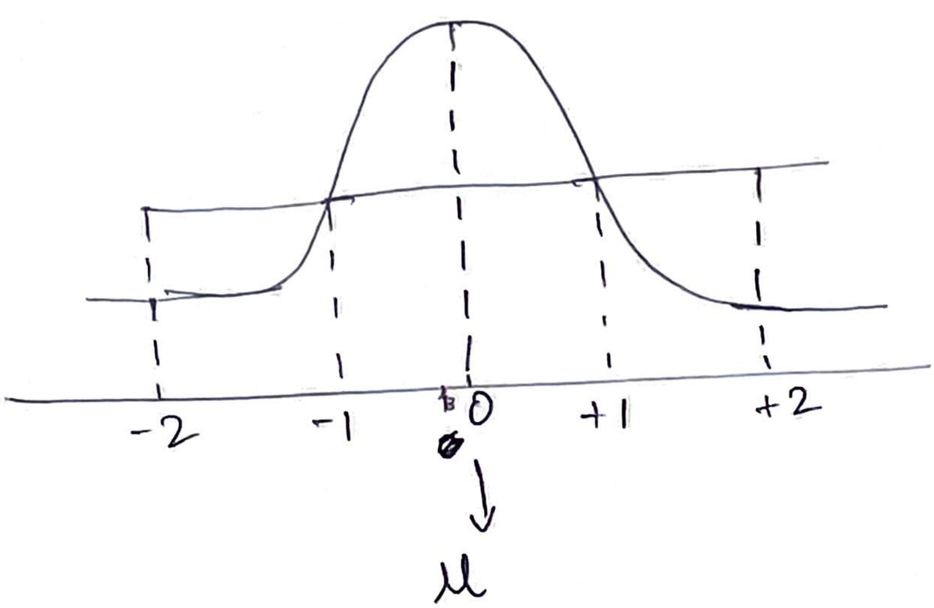


Z Scores :-

Standardizing Scores: Shifting our focus to individual scores.

here the z-score is a statistical technique that uses the mean and the standard deviation to transform each score (X value) into a z-score or a standard score.

* The sign (+) or (−) indicates the location above (+) or below (−) the mean
* The z-score itself indicates distance from the mean in terms of the number of standard deviations.



Population z-score :-

$$z = \frac{X - \mu}{\sigma} = \frac{\text{deviation score}}{\text{standard deviation}}$$

Sample z-scores:

$$z = \frac{x - \bar{x}}{s}$$

what effects does this z- score transformation have on the original distributions?

* <u>Shape</u> : Stays the same.
* <u>Mean</u> : z- score distribution mean is always = 0
* <u>Standard Deviation:</u> is always 1 in z-score distribution

Basically just relabelling the x- axis.