* Inferential statistics :-

Distribution :- 3 types: sampling, central limit theorem, confidence interval.

- sampling distribution is a type of probability distribution created by drawing many random samples of a given size from population.
- Distribution is simply collection of data. or scores or variables orderd from smallest to largest & can be presented graphically

standard Error :- It measures accuracy with which a sample distribution represents a population by using std. deviation.

- sample mean deviates from actual mean of a population. this deviation is std error of mean.

$$SE = \frac{\sigma}{\sqrt{n}}$$

SE = std error
σ = sample std deviation
n = number of sample

- It indicates how different population mean is likely to be from sample mean.

Estimators & Estimates :-

- Estimator is a statistic that estimates some fact about population. It is a rule that creates an estimate.

  e.g. sample mean($\bar{x}$) is estimator of populatⁿ mean μ

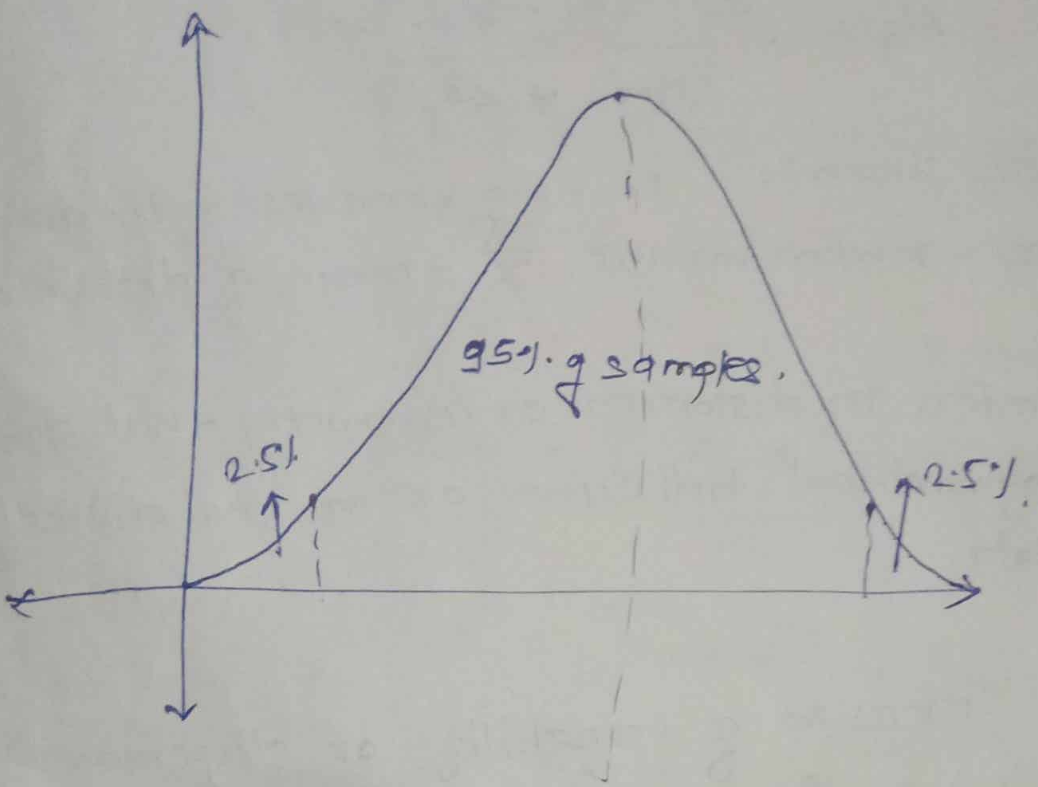
- Estimator is a fnⁿ of sample & estimate is a value of an estimator calculated from a sample.

* central Limit Theorem :-

- Average of your sample means will be the population mean.

- Add up the means from all of your samples, find the average & that average will be your actual population mean.

- confidence interval :-



$$CI = \bar{x} + z \frac{s}{\sqrt{n}}$$

CI - confidence interval

$\bar{x}$ - sample mean

z - confidence level value

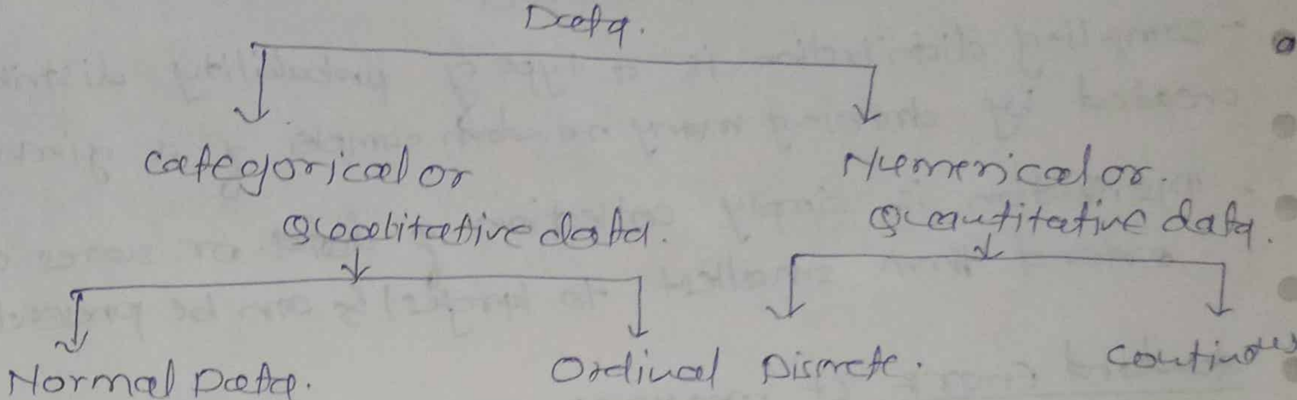
s - sample std. deviatⁿ

n - sample size

- CI is range of values that we observe in our sample & for which we expect to find the value that accurately reflects the population.

* Types q data:-

- Hominal.
- ordinal
- Discrete
- continoues



e.g. gender, home town, colour etc.
Birthdate, favarite sport, school postcode.
↓
Numeric values but dont have numerical meaning.

- Hominal:- Label variables without providing numerical value. Also called nominal scale. It cannot be ordered or measured. e.g. Letters, symbols, words, gender etc.

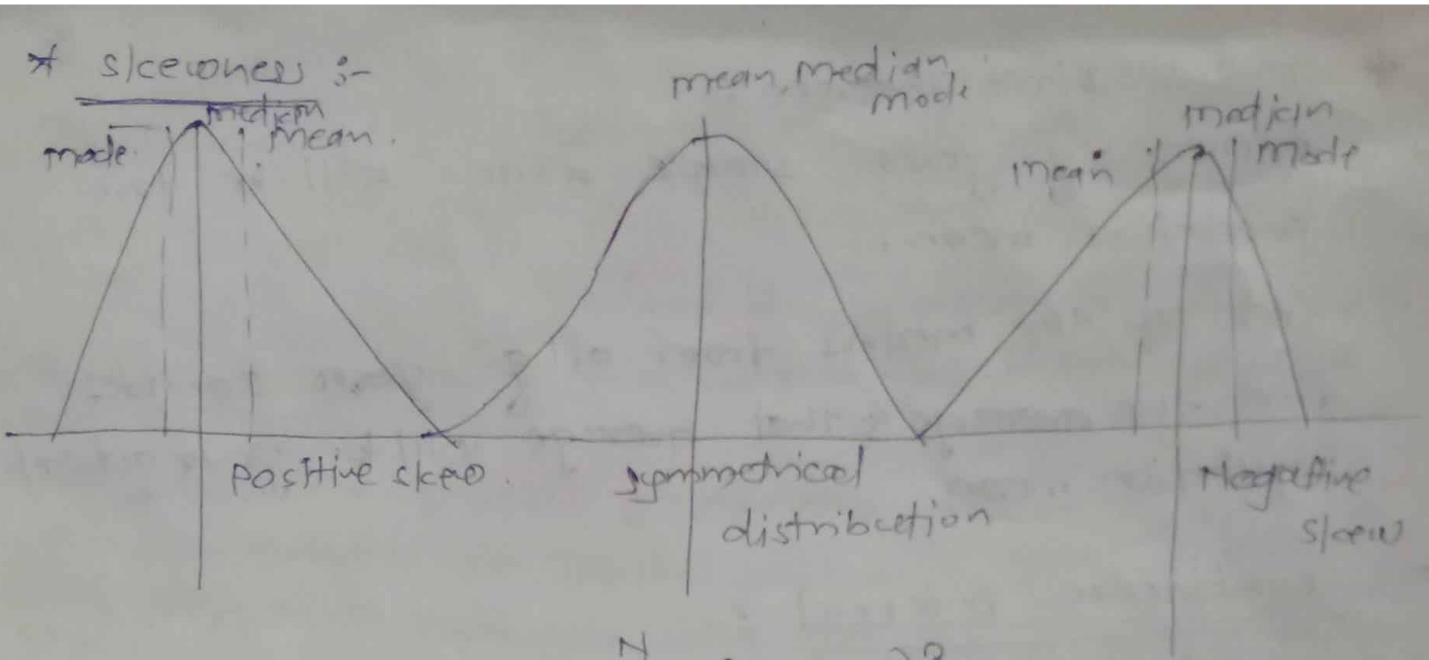
- Ordinal:- It follows natural order Represented using bar chart found in surveys, finance, economics, questionare etc

- Discrete data:- Takes only discrete values. contains only finite no q possible values. Things can be counted in whole numbers & cannot be subdivided meaningfully.

e.g. no q students in the class.

- continoues data:- It can be calculated & has an infinite no q probable values that can be selected within a given specific range.

* skewness :-



$$\tilde{\mu}_3 = \frac{\sum_i^N (x_i - \bar{x})^3}{(N-1) * \sigma^3}$$

μ_3 - skewness N - no. of variables in the distribution
x_i - random variable $\bar{x}$ - mean of distribution

- skewness refers to distortion or assymetry that deviates from the symmetrical bell curve or normal distribution in a set of data.

* variance :- Measure of variability or dispersion

- It measures how far a set of numbers is spread out from their average value.

- measurement of spread bet^n numbers in a data set
- Measure how far data pts from their mean.

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

S^2 = sample variance.
x_i = value of one observation.
$\bar{x}$ = mean value of all observations.
n = no. of observations.

10

II Descriptive & Inferential statistics

* Descriptive statistics :-

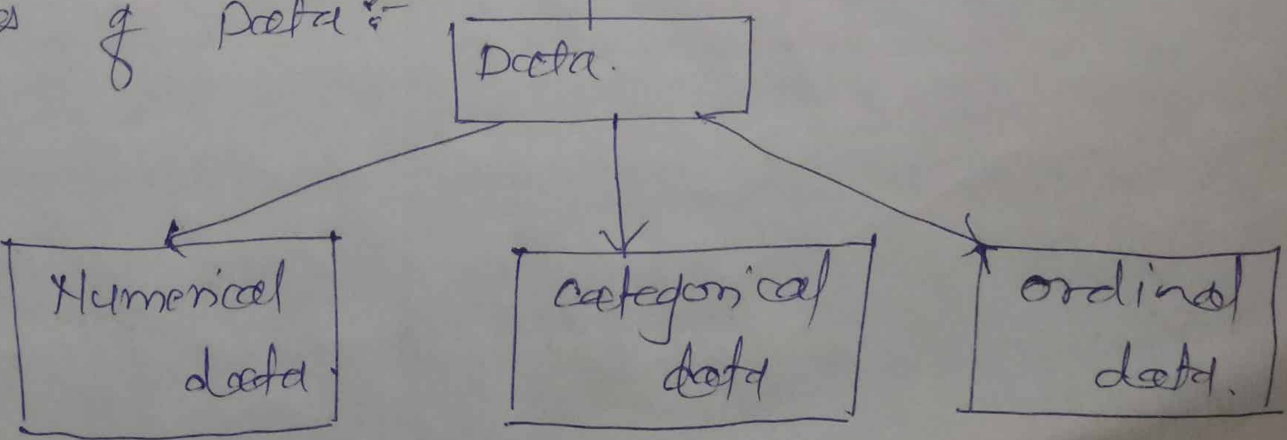
1. Population & sample :-

- Population is the entire group that you want to draw conclusions about.

- A sample is the specific group that you will collect data from. size of the sample is always less than the total size of the population.

Population	sample.
1. Advertisements of IT jobs in Netherland.	Top 20 search results for advertisements for IT jobs in Netherlands on June 1 2021
2. songs from Eurovision song contest.	Winning songs from Eurovision song contest that were performed in English.
3. Undergraduate students in Netherlands	300 ug students from 3 universities for research study.

* Types of Data :-



* Numerical data :-

- Data includes count or measurement of any object or person such as mass, volume, height, sugar level etc.

* categorical data :-

- The characteristics or behavioural attribute of person or object is included under categorical data. It can be in any form such as sugar, gender, nativity, caste, marital status or a favorite thing.
- It can be either true or false.

* ordinal data :-

- Mixture of numerical & categorical data.
- Data is arranged under categories & numbers are placed in the categories which give a correct meaning.

e.g. Hotel rating can be given from zero to five with the category from poor to excellent considering its ambience, taste, service, cost facilities etc. By merging data, the chart is

- prepared to analyze the performance of the hotel.

(2) * Measurement levels :-

1 Nominal	2 ordinal	3 Interval	4 Ratio
- categorized.	categorized & ranked data.	categorized ranked & evenly spaced.	—||— & has a natural zero.

1. Nominal level :-

- You can categorize your data by labelling them in mutually exclusive groups, but there is no order bet^n the categories.

ordinal Level :-

You can categorize & rank your data in an order but you cannot say anything about the intervals bet^n the rankings.

e.g. You can rank top-5 olypic medalists. this scale does not tell you how close or far apart they are in number of wins.

Interval level :-

You can categorize, rank & infer equal intervals bet^n neighbouring data pts, but there is no true zero point.

The diff. bet^n any two adjacent temp. is the same: one degree. But zero degrees is defined differently depending on the scale.

* Representation of categorical variables :-

- categorical data is the statistical data type consisting of variables or of data that has been converted into that form, for example as grouped data.

- More specifically, categorical data may derive from observations made of qualitative data that are summarised as counts or cross tabulations. or from observations of quantitative data grouped within given intervals. often, purely categorical data are summarized in the form of a contigency table. However, particularly when considering data analysis, it is common to use the term 'categorical data' to apply to data sets that, while containing some categorical variables, may also contain non-categorical variables.

- A categorical variable that can take on exactly two values is termed a binary variable or dichotomous variable.

- categorical variables with more than possible values are called polytomous variables : categorical variables are often.

* Dummy coding
The reference group is assigned a value of 0 for each code variable

* Effects coding
- Data are analyzed through comparing one group to all others. groups.

* contrast coding :-

- It allows researcher to directly ask questions
- ~~This tailored. hypothesis~~
- sum of contrast codes is restricted by three rules:
- sum of contrast coefficients per each code variable must equal zero.
- Diff. betn sum of positive coefficients & the sum of the negative coefficients should equal 1.
- Coded variables should be orthogonal.

* Nonsense coding :-

- It occurs when one uses arbitrary values in place of designated 0's 1's & -1' seen in previous coding systems.

* Embeddings :- Embeddings are codings of categorical values into high-dimensional real valued vector spaces, usually in such a way that similar values are assigned similar vectors or with respect to some other kind of criterion making the vectors useful for resp. applu.

⑨ * Measures of central tendency:-

* Mean :- The most commonly used measure of central tendency is the mean. To calculate the mean of a dataset, you simply add up all of the individual values & divide by the total no. of values

Mean = (sum of all values) / (total ~~no~~ of values)

* Median :- The median is the middle values in a dataset. Arrange all values in a dataset from smallest to largest & finding the middle value will give you median. If there are odd no. of values middle one is median. & If there are even no. of values the median is the average of the two middle values.

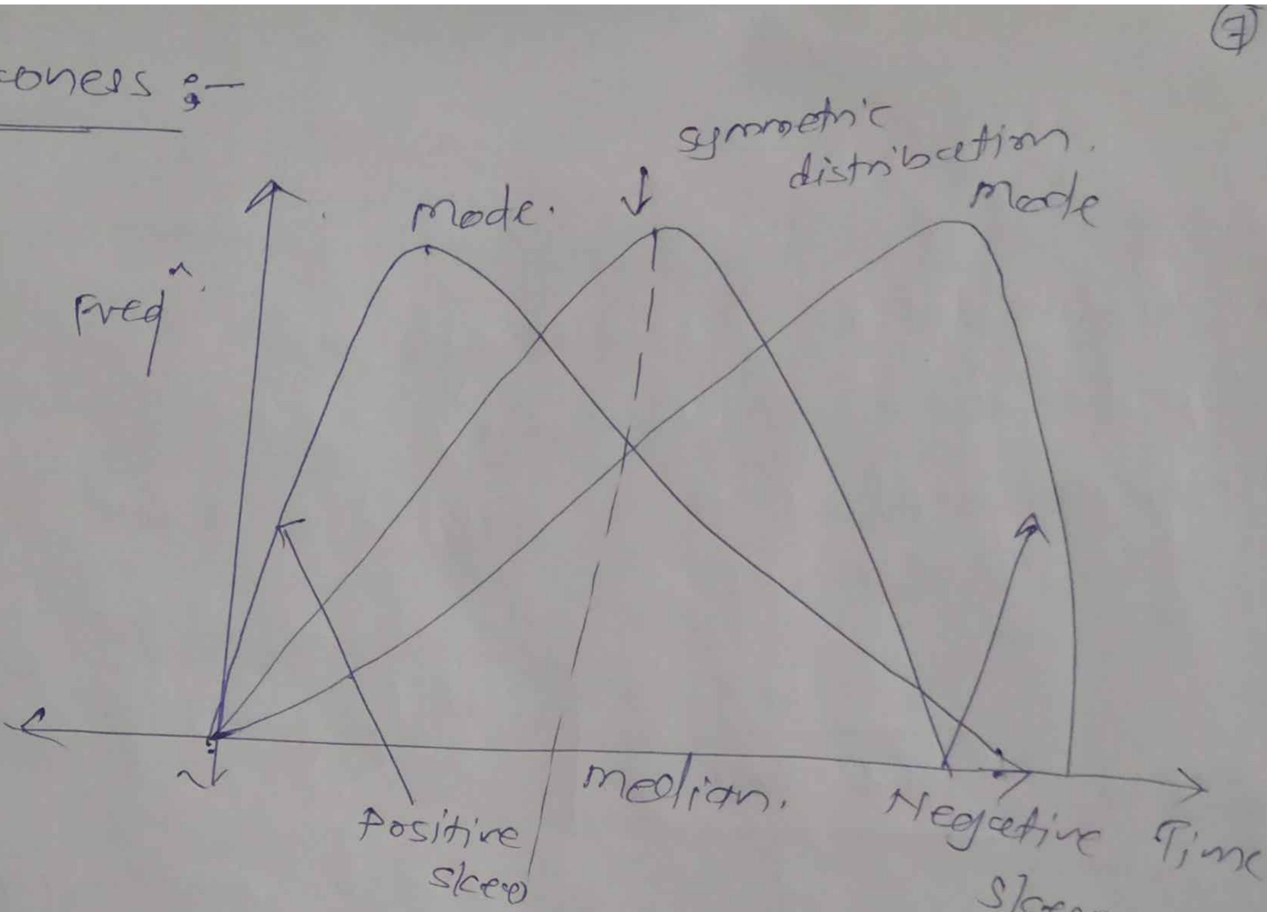
e.g. Player.	1	2	3	4	5	6
Home Runs.	8	9	10	12	13	14

$$= \frac{10+12}{2} = \frac{22}{2} = \underline{\underline{11}}$$

* Mode :- Mode is the value that occurs most often in a dataset. Dataset can have no mode one mode or multiple modes.

e.g.	1	2	3	4	5	Mode - 10
	8	9	10	10	11	is

* Skewness :-



* Positive skewness

- If the given distribution is shifted to the left with tail on the right side, it is a positively skewed distribution.
- It is also called as right skewed distribution.
- It assumes skewness value of more than zero.
- mean value is greater than the median & moves towards right. ~~mean~~ mode value occurs at highest freq^n of distribution

Negative skewness.

- If the given distribution is shifted to the right & with the tail on the left side, it is a negatively skewed distribution. It is also called a left-skewed distribution.
- The skewness value of any distribution showing a negative skew is always less than zero.
- $$\text{skewness} = \frac{\bar{x} - M_o}{s}$$

x = mean value M_o = Mode value
s = standard deviation of sample data.

* Variance :- Is the expected value of the squared variation of arandom variable from its mean value in probability & statistics. Informally, varaince estimates how far a set of numbers are spread out from their mean value.

$$var(x) = E[(x) - \mu)^2]$$

* standard Deviation :- It measures dispersion of a dataset relative to its mean & is calculated as the square root of variance.

$$\text{standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

where, x_i = value of the i^{th} pt in the dataset
$\bar{x}$ = The mean value of the dataset
n = The number of data points in the dat

* variance vs std. deviation :-

Parameters	variance	std. deviation
Meaning	Numerical value that describes variability of observations from its arithmetic mean.	—||— measure of dispersion of observations within dataset
What is it?	It's average of squared deviations.	It is root mean square deviation.
Labelled as	sigma-squared (σ^2)	sigma (σ).
Expressed in	squared units	same units as the values in the set of data.
Indicates	How far individuals in a group are spread out.	How much observations of a data set differs from its mean.

* coefficient of variation :-

It is statistical measure of dispersion of data pts around the mean.

The metric is commonly used to compare the data dispersion betn distinct series of data.

By ~~identifying~~ determining coefficient of variation of different securities, an investor identifies the risk to reward ratio of each security & develops an investment decision.

$$\text{coefficient of variation} = \frac{\sigma}{\mu} \times 100\%$$

σ - standard deviation μ - mean

e.g. stocks:

$$\text{coefficient of variation} = \frac{\text{volatility}}{\text{Expected Return}} \times 100\%$$

Fred was offered stock of ABC corp. It is a mature company with strong operation & financial performance. The volatility stock is 10% & the expected return is 14%.

* covariance & correlation :-

Parameters	Covariance	Correlation
Meaning.	It indicates extent of variable being dependent on each other. Higher values denotes higher dependancy.	—||— signifies strength of association bet^n the variables when other things are constant.
Relationship.	—||— correlation can be gathered from covariance.	correlation gives value of covariance on a std. scale.
Values	lie bet^n -∞ to +∞	correlation has limited values in range of -1 & +1
scalability	Affects covariance	correlation isn't affected by a change in scale.
Units.	covariance will have a definite unit as it is concluded from multiplication of numbers & their units	correlation is a number without units but includes decimal values.

9096116898.

* coefficient of variation :-

$$CV = \frac{\sigma}{\mu}$$

σ - populatⁿ std dev
μ - ||- mean.

- Ratio of std. deviation to mean higher the coeff. of variation, greater the level of dispersion around the mean.
- Used to determine variability of data.

* Find the coeff. of variatⁿ of following sample:

$\{1, 5, 6, 8, 10, 40, 65, 88\}$.

Ans:- Mean = $1+5+6+8+10+40+65+88/8 = 223/8$

$\mu = 27.875$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (1 - 27.875)^2 = 7578.875$$

$$\text{Variance} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{7578.875}{8-1} = 1082.696$$

$$\text{std deviat}^n = \sqrt{\text{variance}} = \sqrt{1082.696} = 32.904$$

$$CV = \frac{\sigma}{\mu} = \frac{32.901}{27.875} = 1.180.$$

---

* Std. deviation:- $\sigma = \sqrt{\Sigma(x_i - \mu)^2 / N}$.

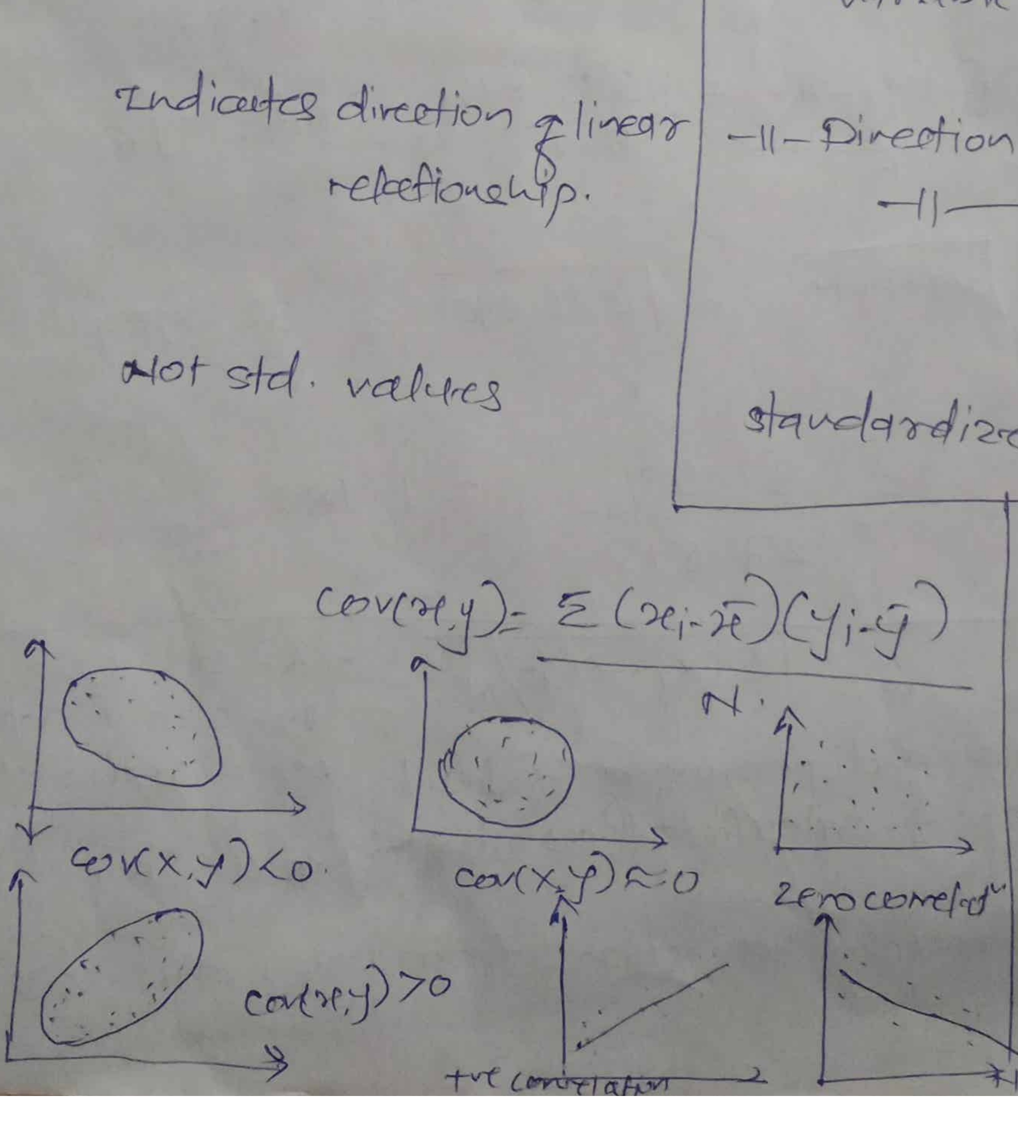
- σ - std. dev.
N - size of populatⁿ
x_i - each value of pop
μ - populatⁿ mea

It is the measures of dispersion of dataset relative to its mean & calculated as square root of variance.

It shows how much variation from the mean exists.

key Parameters	covariance	correlation
meaning	Indicates the extend to which two random variables change in random	statistical measure that indicates how strongly two variables are releted.
What it is	Measure of correlation	scaled version of covariance
Values range	$-\infty$ to $+\infty$	-1 to $+1$.
change in scale	Affects covariance	Does not affect correlation.
	Affected by scale of variable.	Not affected by scale of variable
	Indicates direction of linear releationship.	–||– Direction & strength –||——
	Not std. values	standardized values.

$$cov(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$



Covariance indicates how two variables are releted to one another.
covariance shows how two variables differ, & correlation shows you how two variables are releted.