**What is Statistics?**

Statistics is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them. In general, its investigations and analyses fall into two **broad categories called descriptive and inferential statistics**

Maths vs statistics

Maths we get exact values whereas

Statistics we get estimated values

**Descriptive**

Where deals with data to describe the data or understand the data like from sales data most are from male side or from new nork cities.

Eg mean, median, sd/variance, finding the outliers.

**Inferential**

Where we deal with prediction of data using the techniques

**Population and samples**

Population – refers to total amount of things

Samples – small portion of population that is used for study

**Sample size- total things in sample**

Variable: How a particular person can have different heights, weights and hair colour. These are variables. It refers to characteristics of what we are studying. Can vary among different individuals.

Descriptive Methods

Measure of central tendency- mean, median, mode

Dispersion- SD, VARIANCE, range of IQR(inner quartile range)

Shape-skewness, keratoses

Mean – average of data

Particular column/n number = mean

Mean is not good representation of data by conclusion

When outlier is not present then mean is good

Median – middle team 1,2,3,4,5 for odd

Sort in ascending order

3 is the median for odd numbers

Mean of two middle values like 1,2,3,4,5,6 for even

3+4/2 is the median for even numbers

When outlier(extreme values) is present then median is good

Mode – most frequent values

Repeated numbers like 1,1,2,2,3,5,5, -- 1,2,5 are most frequent



**Dispersion**

**2,3,6,**72,8,34

Range – [max,min]

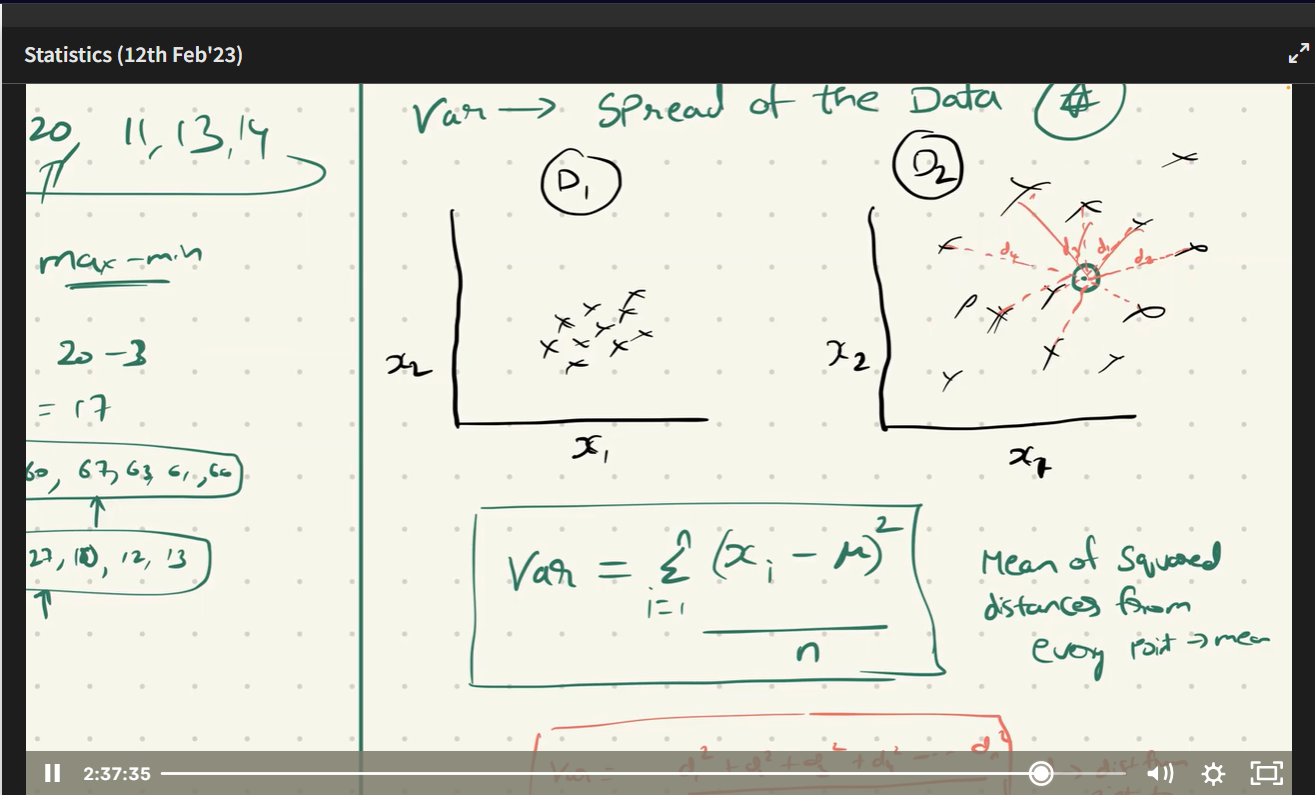
**34-2—max-min that is range**

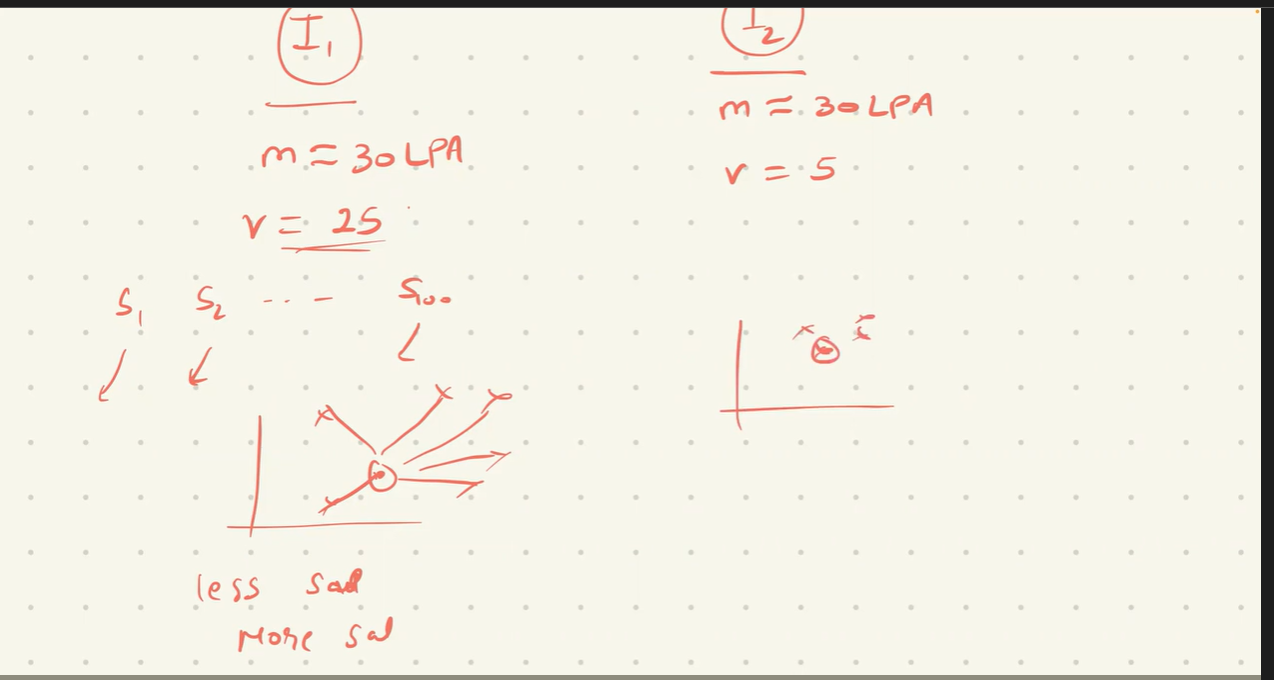
Variance

Variance is to measure the spread of data

How far the data from the mean

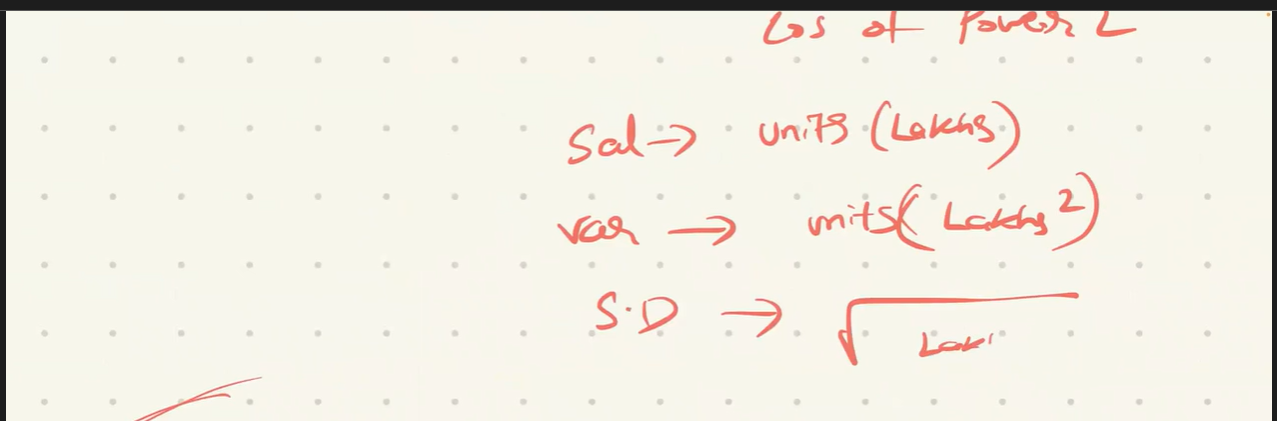
Mean of squared distance from every point





SD- standard deviation

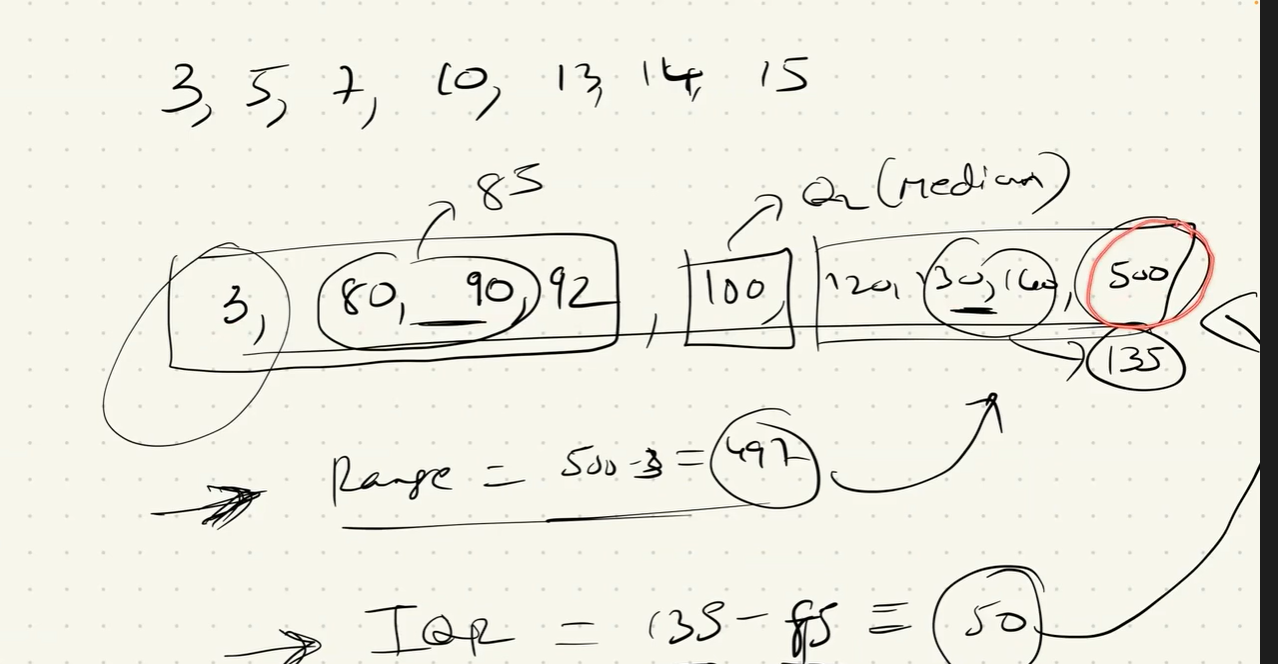
Square root of variance, bcoz SD is better interception of data.

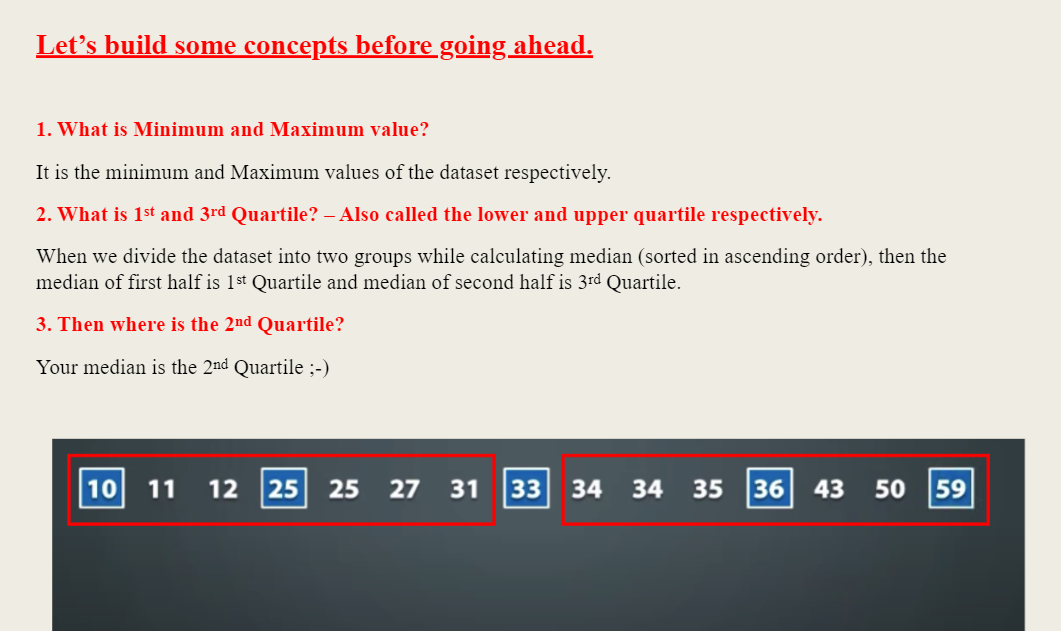


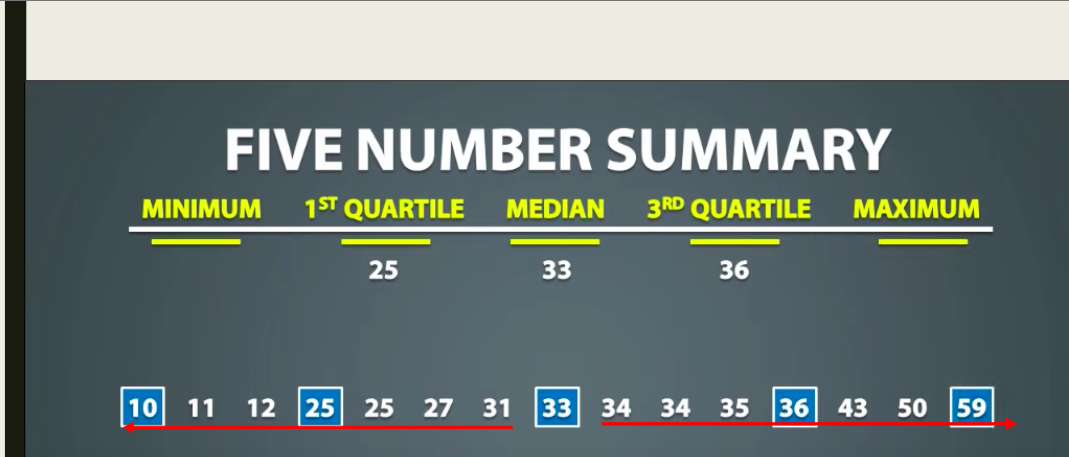
QUANTILE

Q1,q2,q3

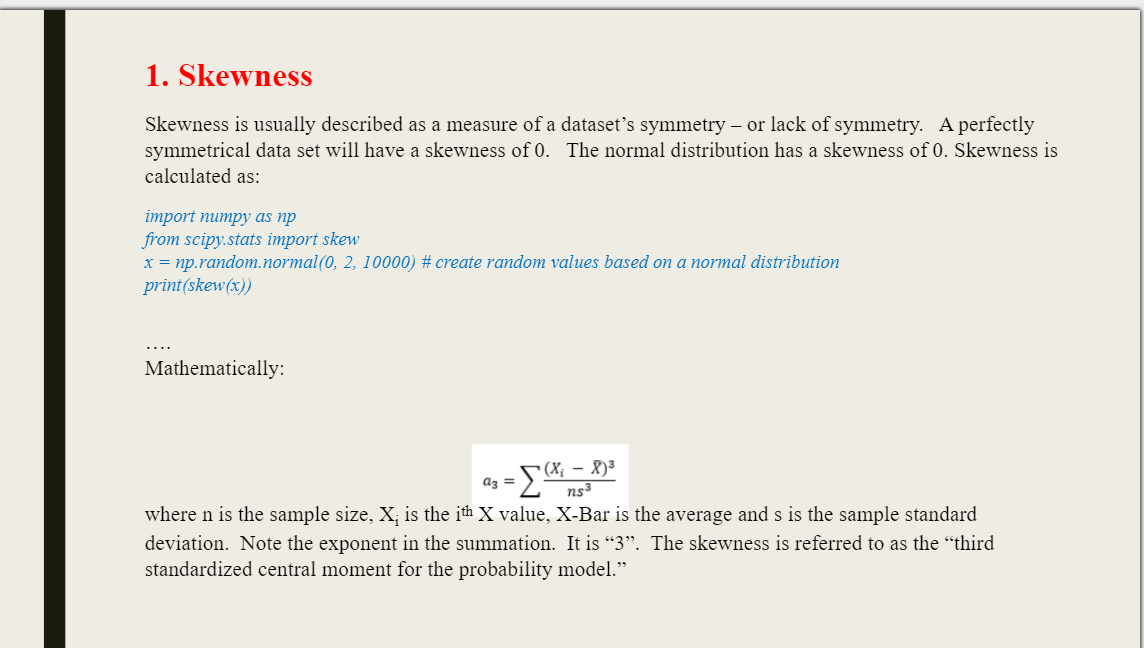
Q1 is left part, q2 is middle , q3 is last part

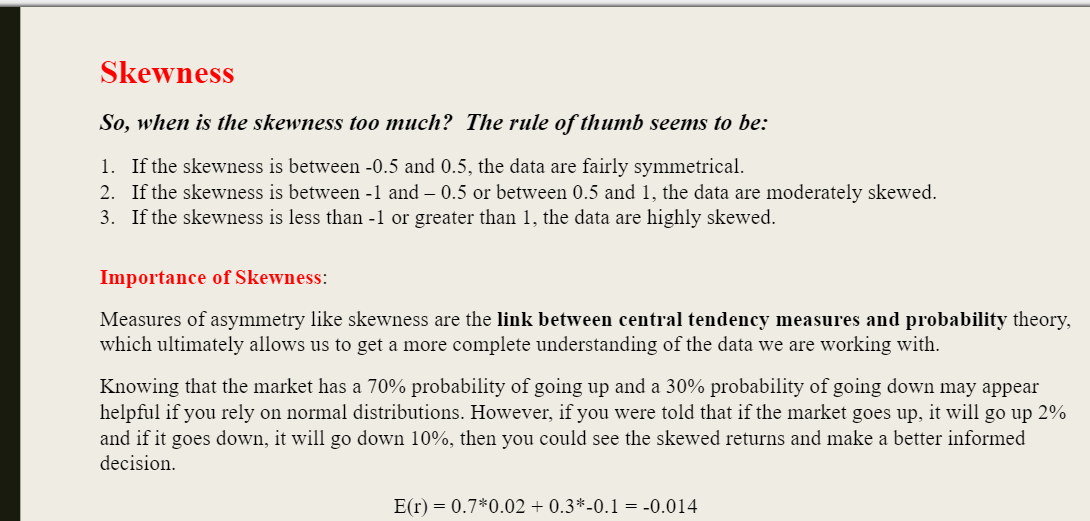




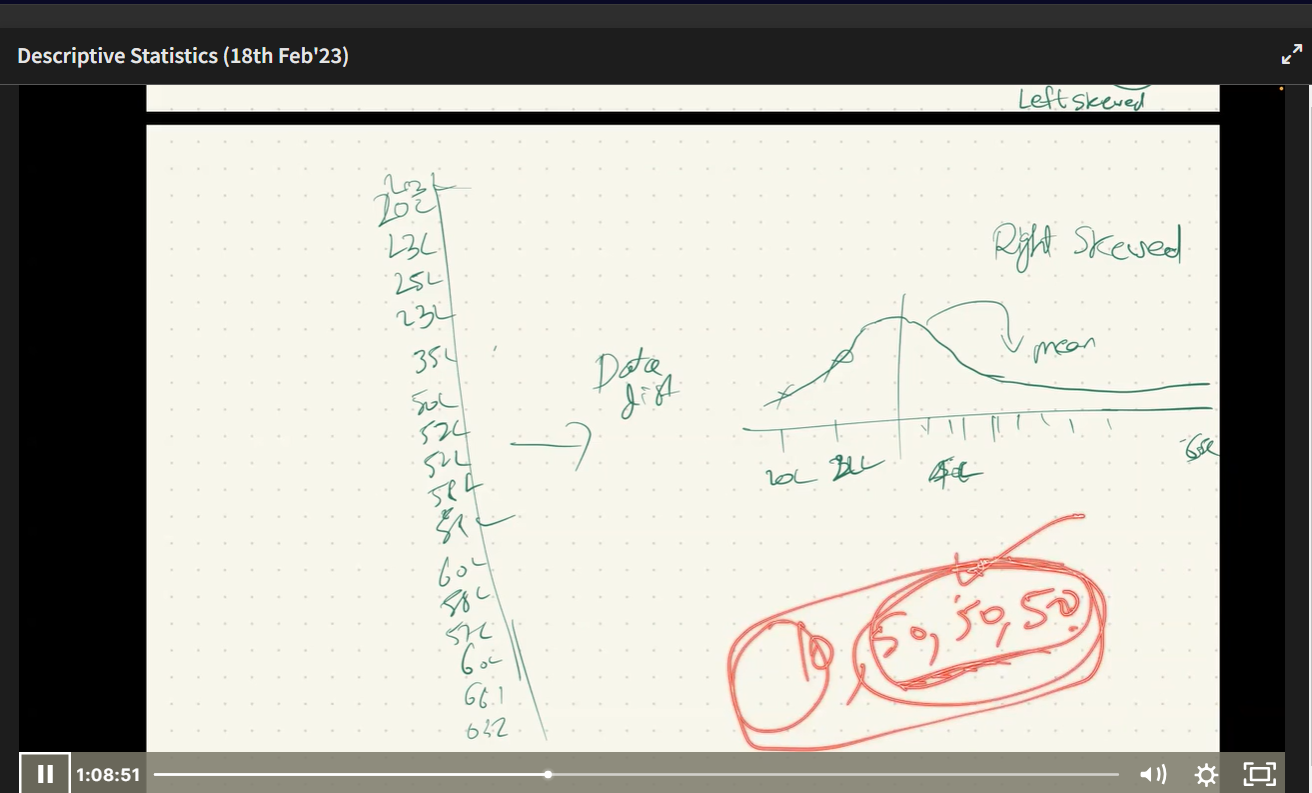


Symmterial is middle which is mean and median in between in normal distribution

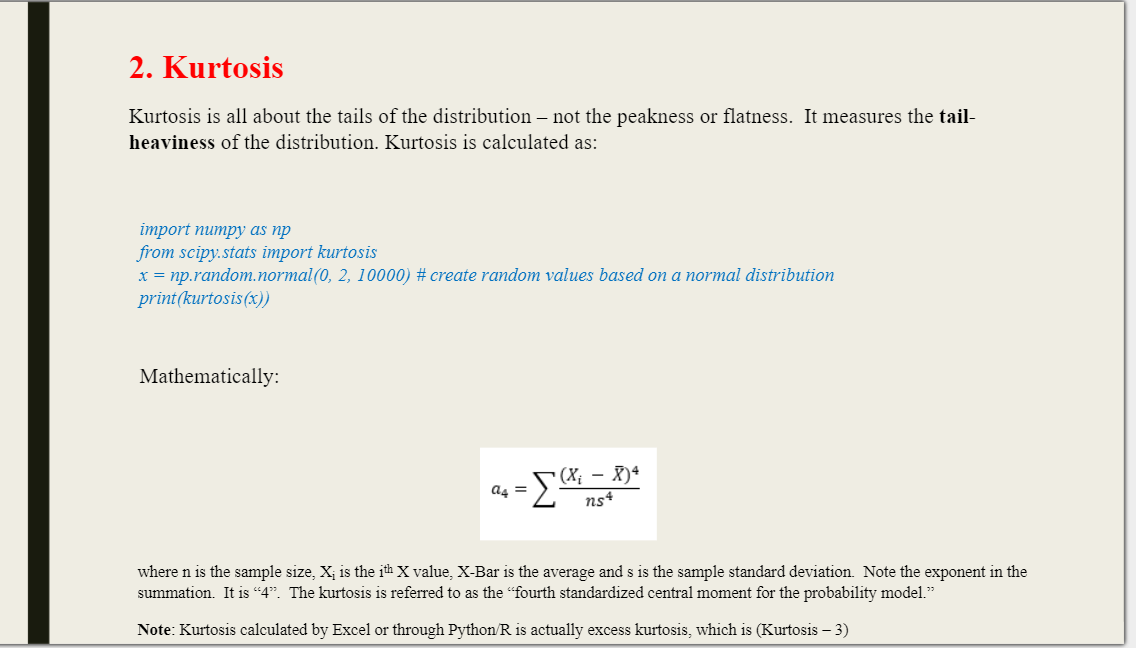


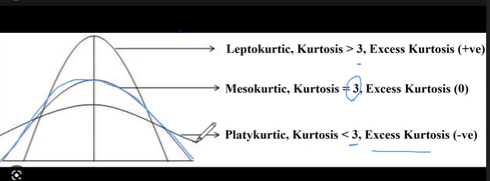


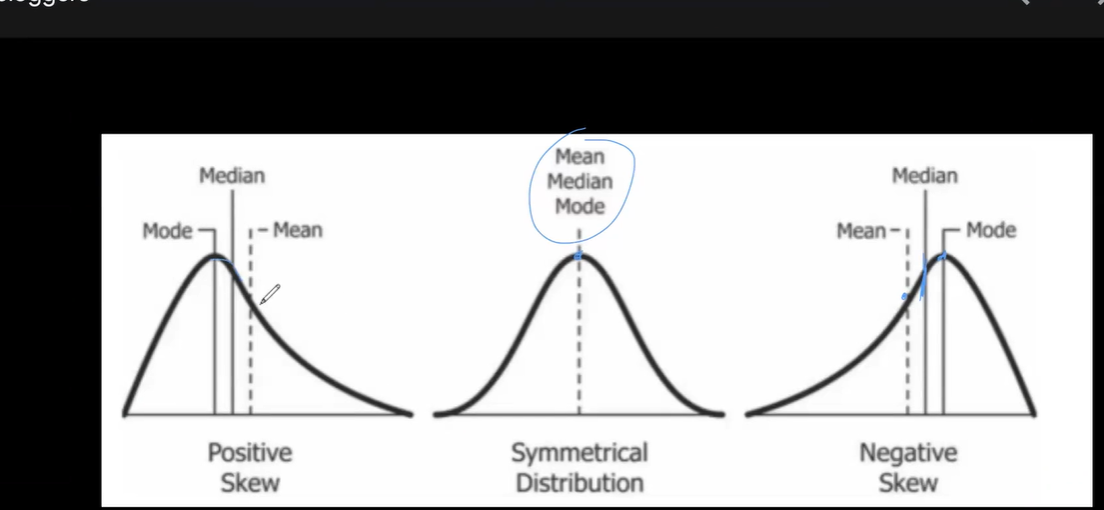
That curve is nothing but the frequency or probability of data



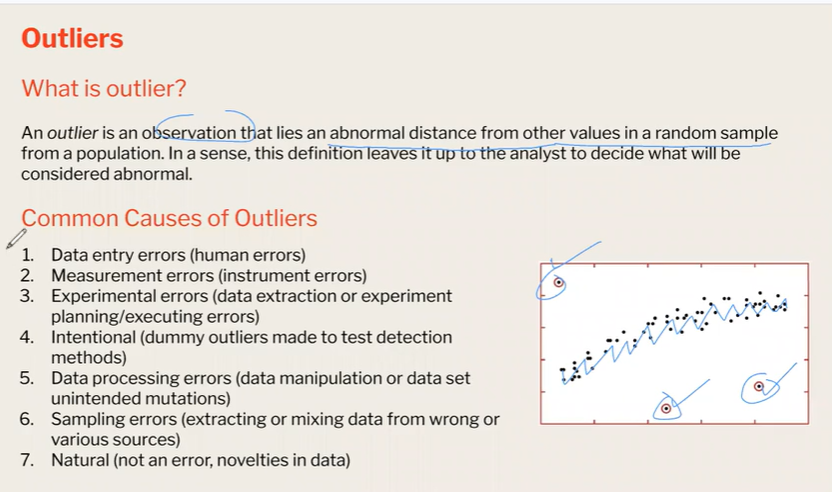
KURTOSIS







Outliers







Normal distribution

It is symmetrical like mean=median=mode

Uniform distribution

For every input data frequeuncy is same

Bernalii distribution

Only two outcomes like coin head or coin

Binomial distribution

Success or failure probability

Population and samples

Mean of samples is apprxomate mean of the population

And it will follow normal distribution

Mean=median=mode

For mormal distribution sample size should be high like 100 ,80

It should not be less size of samples

Then only we get smoothness of curve

**What Is the Central Limit Theorem (CLT)?**

In probability theory, the central limit theorem (CLT) states that the [distribution of a sample](https://www.investopedia.com/terms/s/sampling-distribution.asp) variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

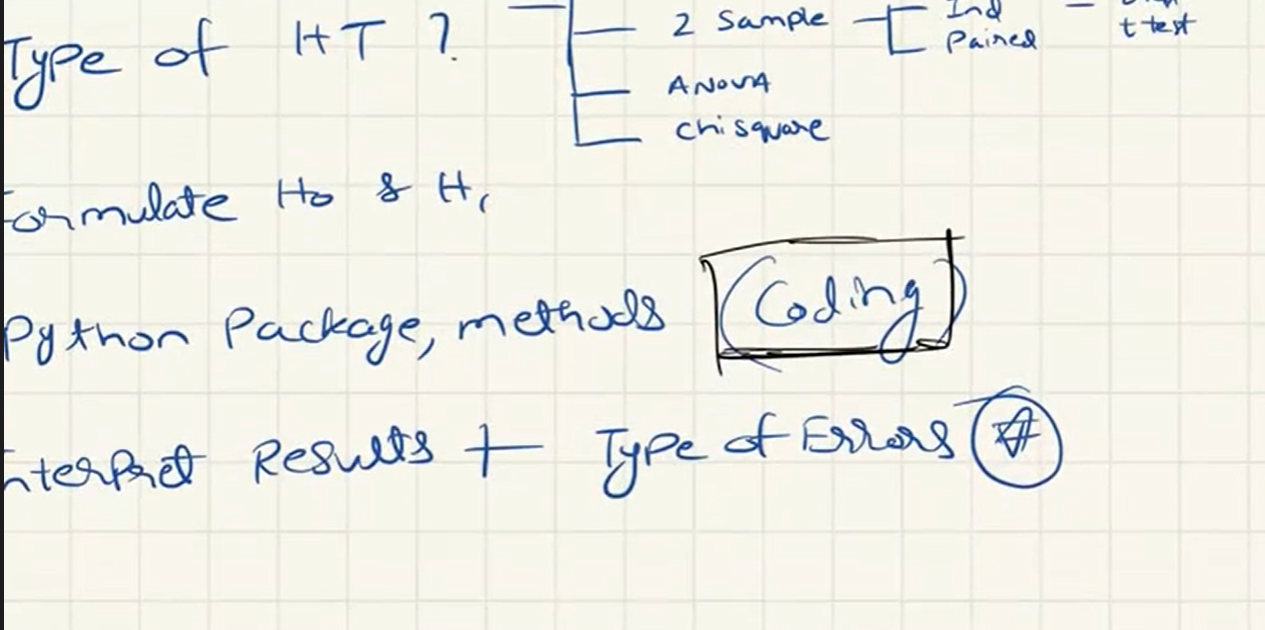
Put another way, CLT is a [statistical](https://www.investopedia.com/terms/s/statistics.asp) premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population. Furthermore, these samples approximate a [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), with their variances being approximately equal to the [variance](https://www.investopedia.com/terms/v/variance.asp) of the population as the sample size gets larger, according to the [law of large numbers](https://www.investopedia.com/terms/l/lawoflargenumbers.asp).

HYPOTHESIS

Type of Hypothesis,-- z\_test,t\_test, anova,chi-square.

Alternative or null hypothesis

Python packages



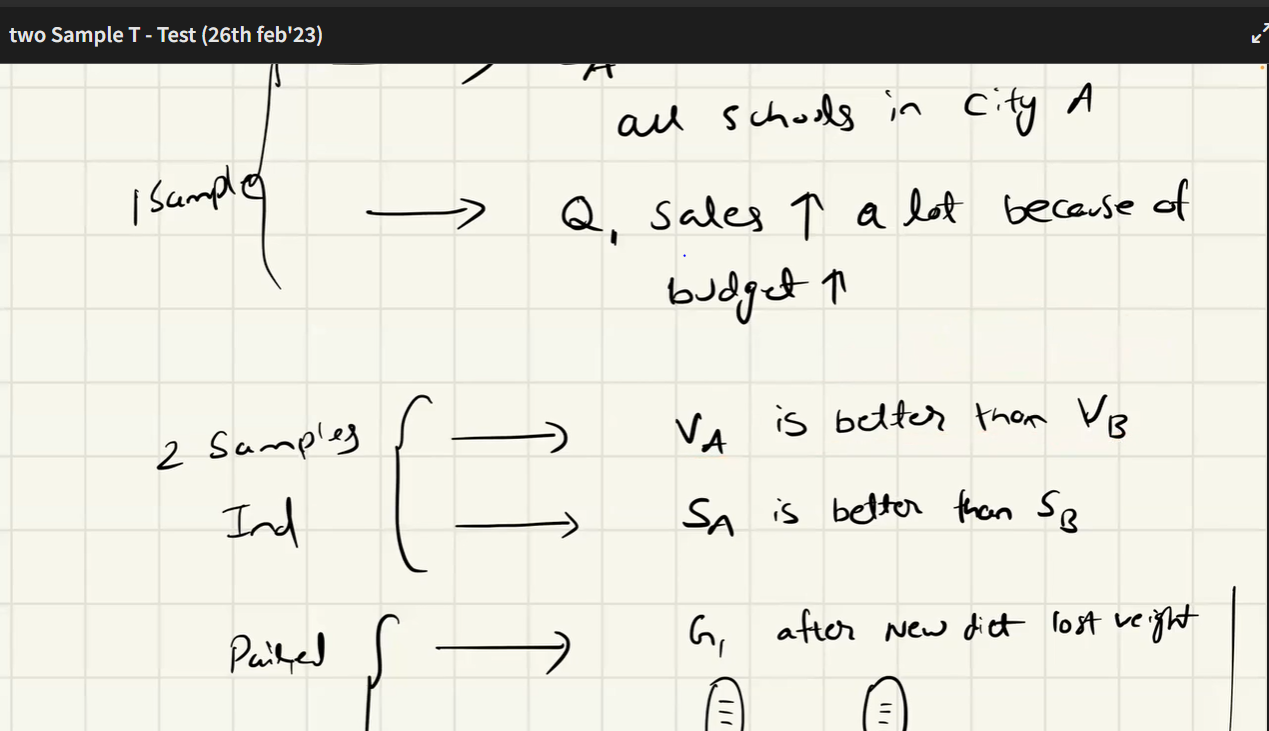
Z-TEST

If samples size n> 30 we go for z\_test when standard deviation of population is known 1sample and 2 samples paired samples

T\_TEST

If sample size n<30 we go for T\_TEST hen standard deviation of population is unknown. 1sample and 2 samples paired samples

Python libraries - scipy



Variance

Spead of data , distance from mean to datapoints in samples

And for only one feature example age

Range of variance is from 0 to infinity

Co-variance

Means between two features like age and salary

Minus infinity to plus inifinity

Correlation

Is the normalised version of co variance



