Statistics 🡪 branch of applied mathematics that involves collection, description, analysis and inference of conclusions from quantitative data.

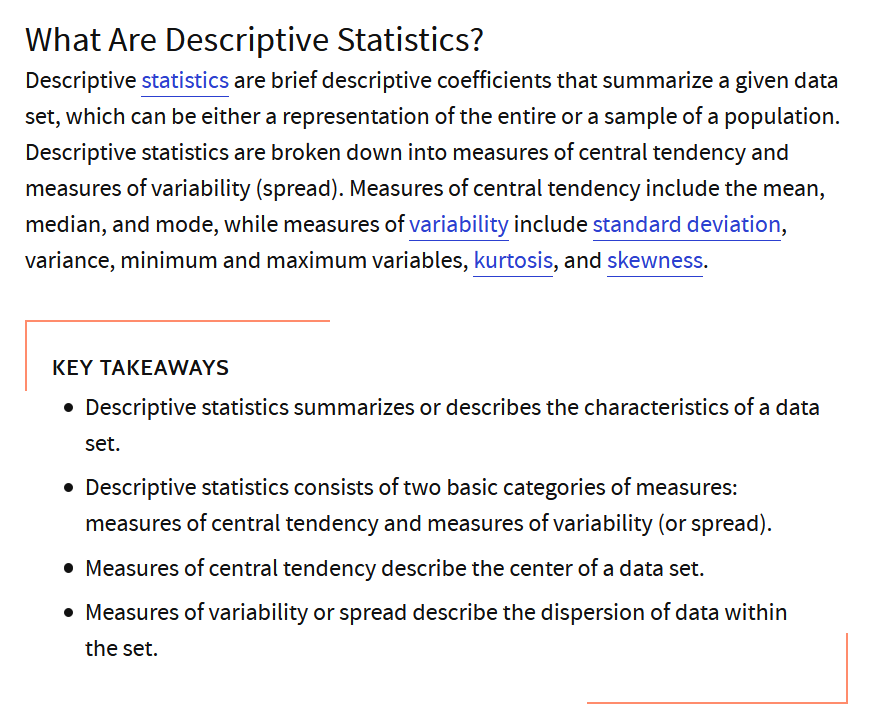
Statistics

Descriptive Statistics Inferential Statistics

## Descriptive Statistics

From the above n age values, what can you do????  
We can find the mean, variance, std, average etc…   
(i.e) We are describing the data using some techniques to find some useful information.

Descriptive 🡪 Describing the data in a proper and to the point. For this description we will apply a technique on sample and population.



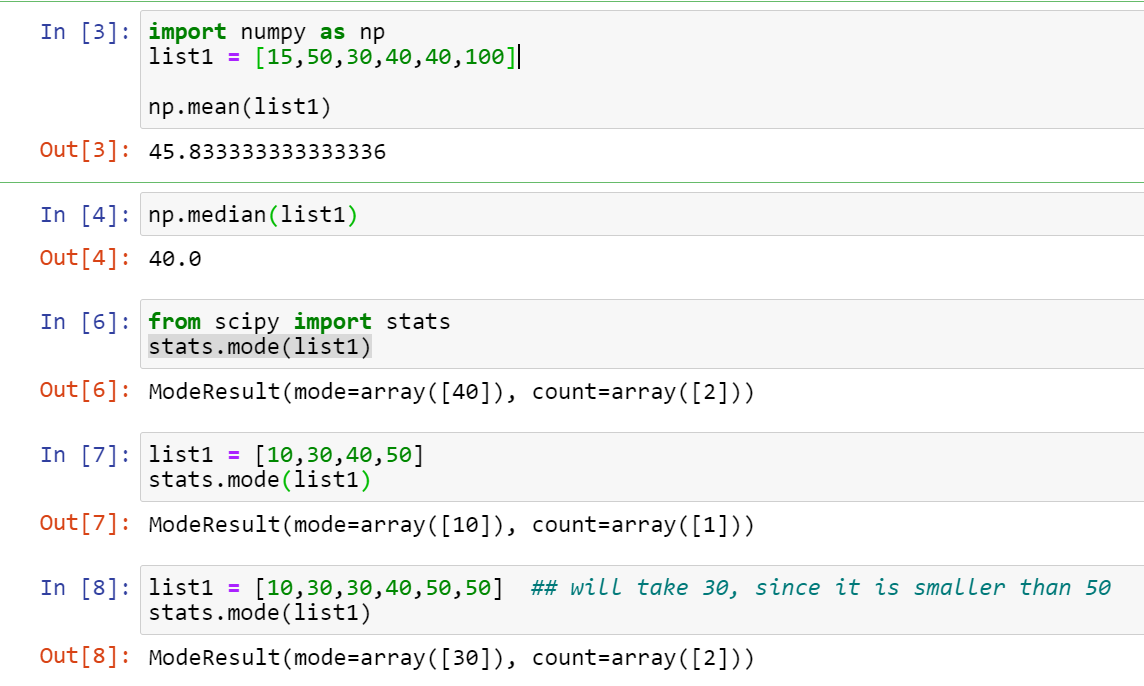
### Population

### Sample

### Measure of Central tendency

1) Mean  
2) Median

To reduce the impact of outlier, median can be used.  
3) Mode  
No need of the data being in ascending order, which-ever having maximum frequency 🡪 mode.

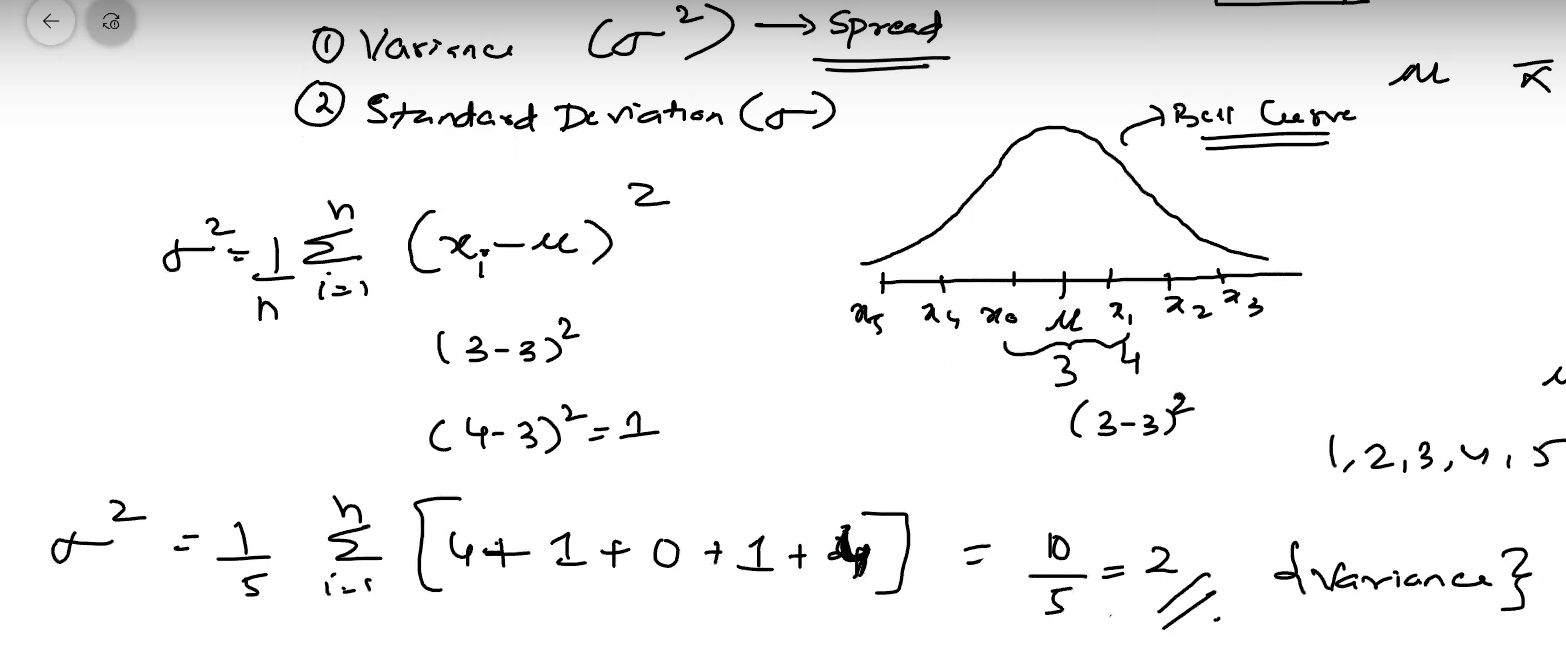


Instead of finding median and mode, remove outliers and find mean, Why can’t we do this??  
Outliers plays a major role, if we remove the outliers then we will be losing the data.

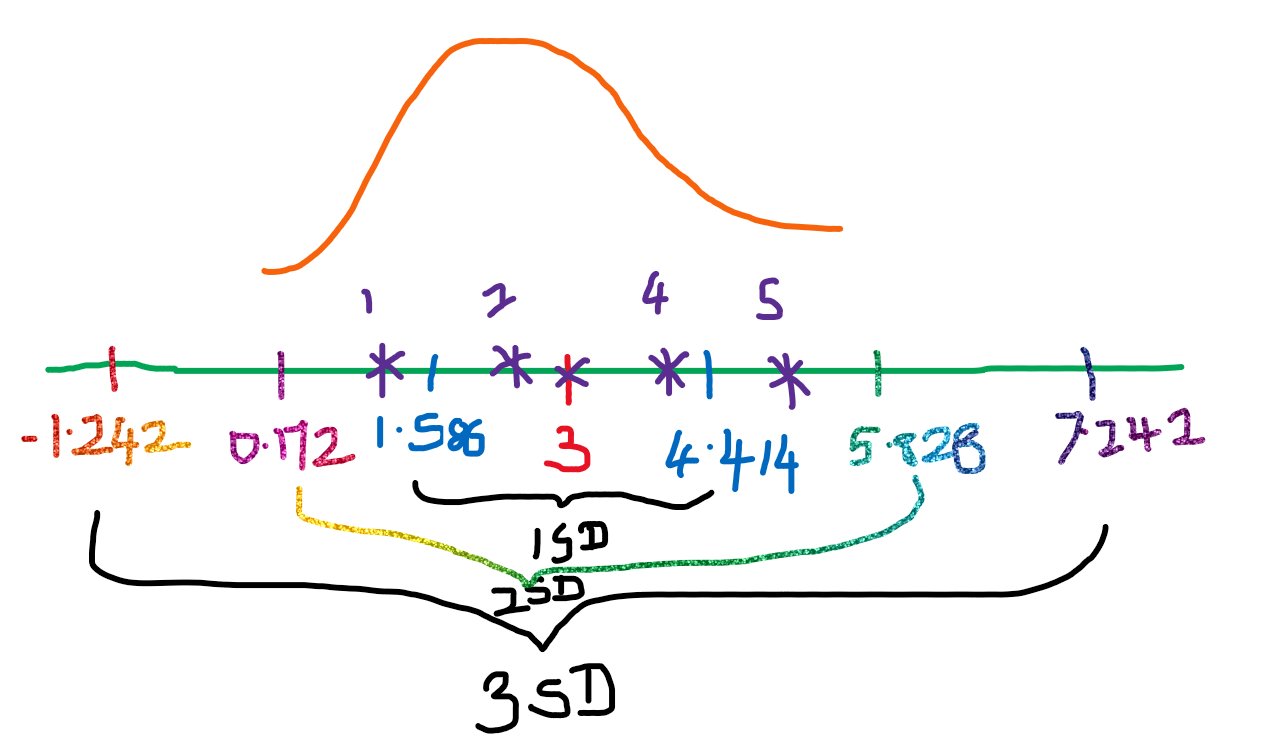
### Measure of Dispersion

## Inferential Statistics

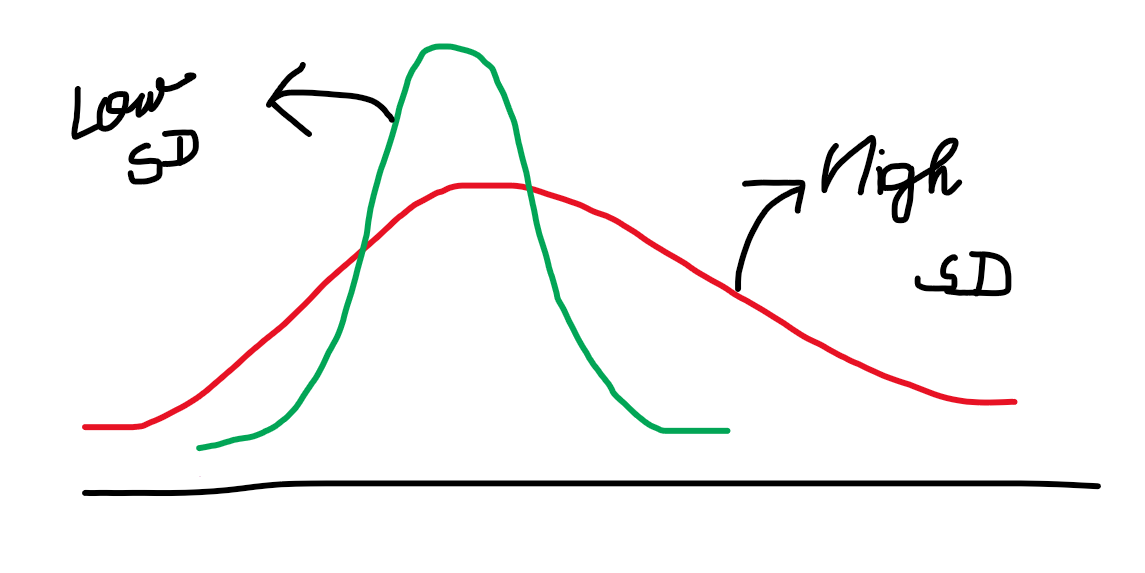
Variance  
Standard Deviation



Difference between Descriptive and Inferential statiscs



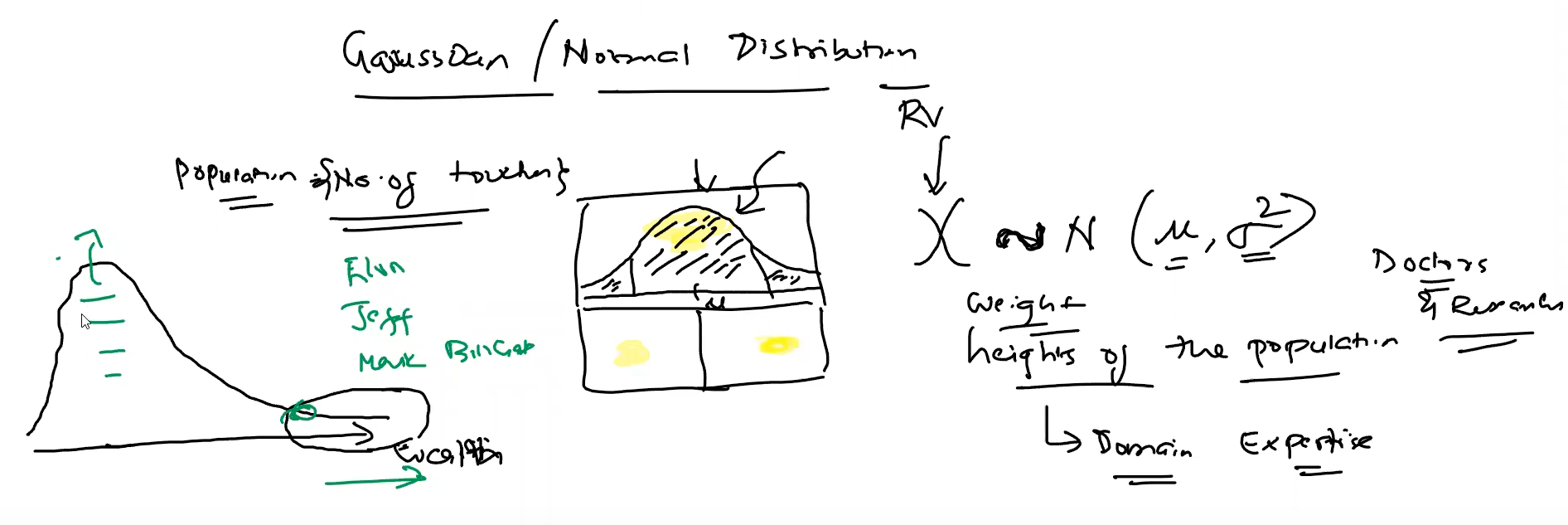
From this we can find spread of the data. How far the elements are from mean.

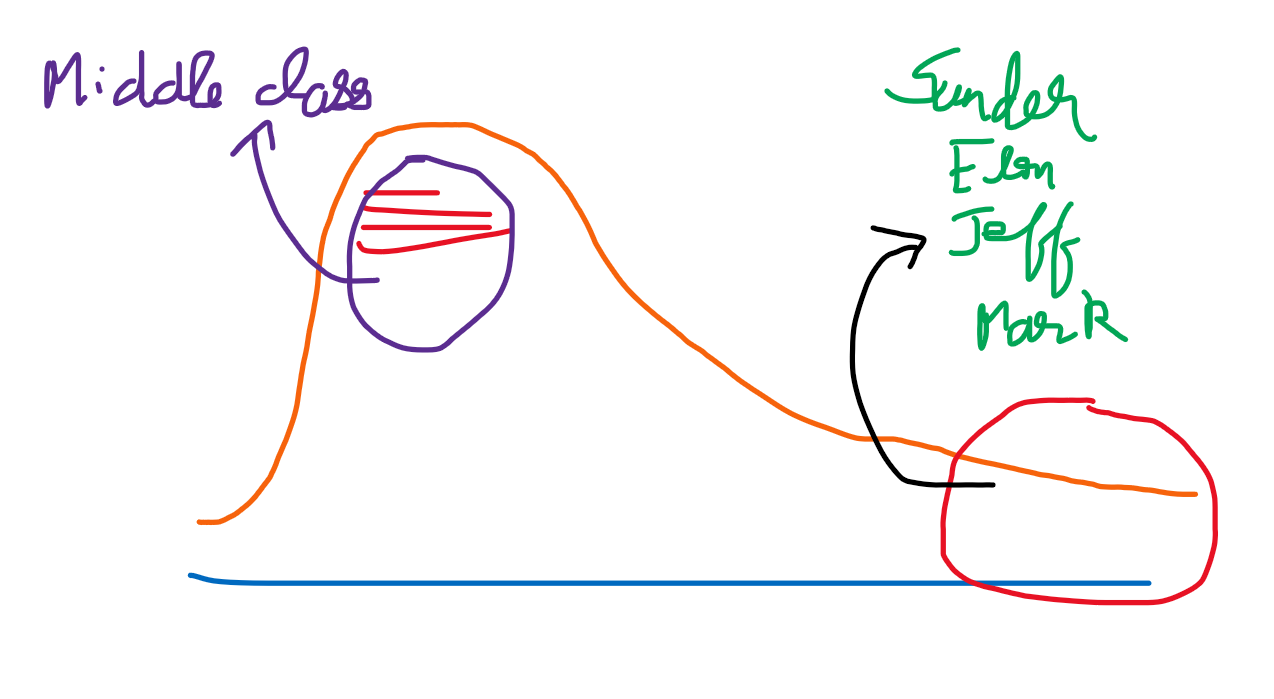


# Random variables

Discrete Random variables  
Continuous Random variables

# Gaussian/Normal Distribution





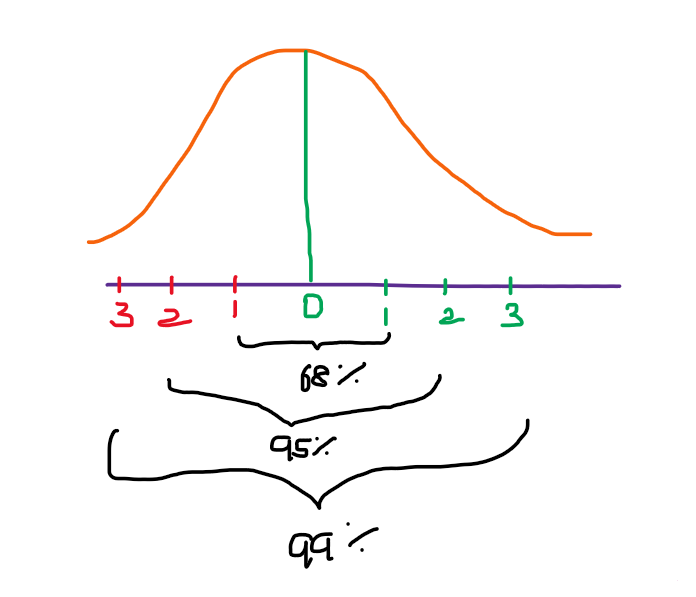
This is Right skewed, and not following Gaussian/Normal distribution

# Standard Normal Distribution

* The standard normal distribution is a normal distribution with a mean of zero and standard deviation of 1.
* The standard normal distribution is centred at zero and the degree to which a given measurement deviates from the mean is given by the standard deviation..
* For the standard normal distribution,   
  68% of the observations lie within 1 standard deviation of the mean;   
  95% lie within two standard deviation of the mean;   
  and 99.9% lie within 3 standard deviations of the mean.

X ~ Gaussian/Normal Distribution (u,Y ~ Standard Normal Distribution (0,1)

Here u=0 and   
It lies between -3 to +3.

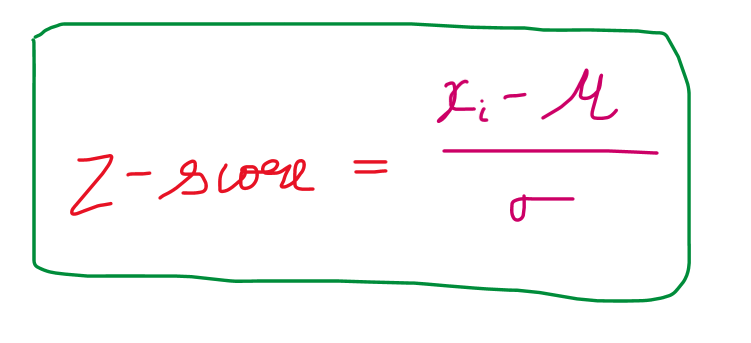


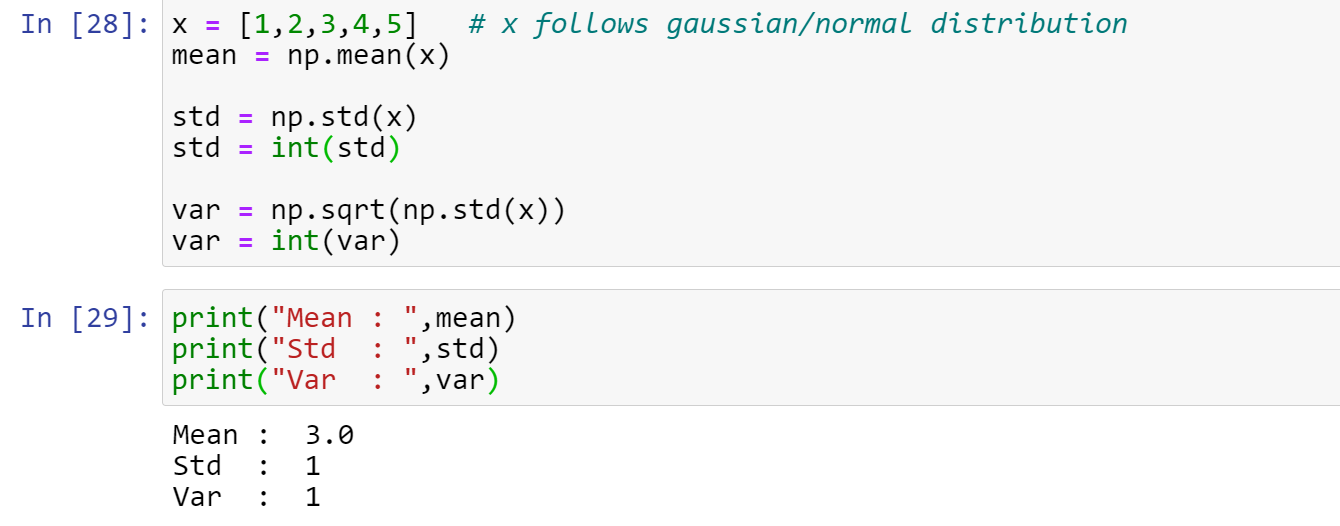
## Converting Gaussian/Normal Distribution to Standard Normal Distribution

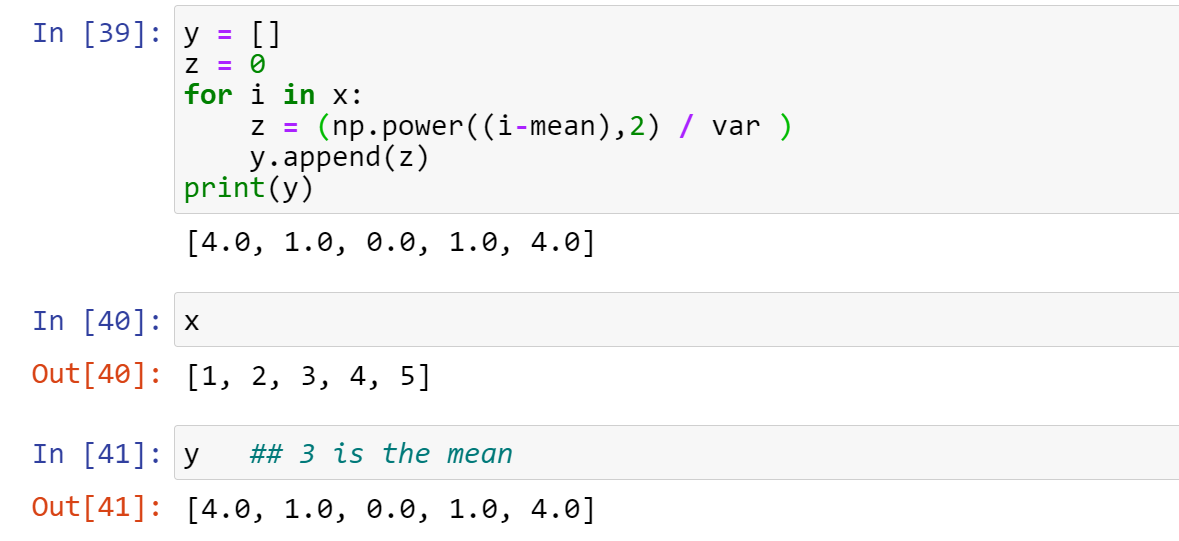
This process is known as Standardization.

What is the purpose of converting ???







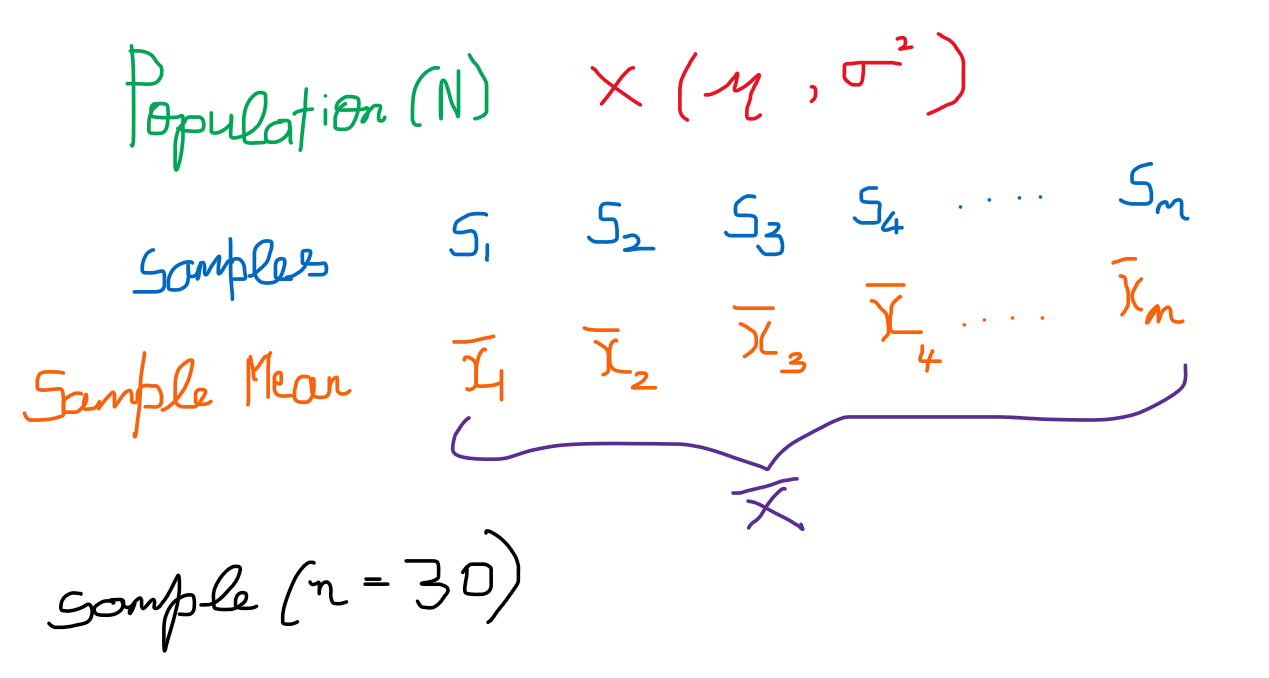


x is converted to y [process 🡪 standardization]

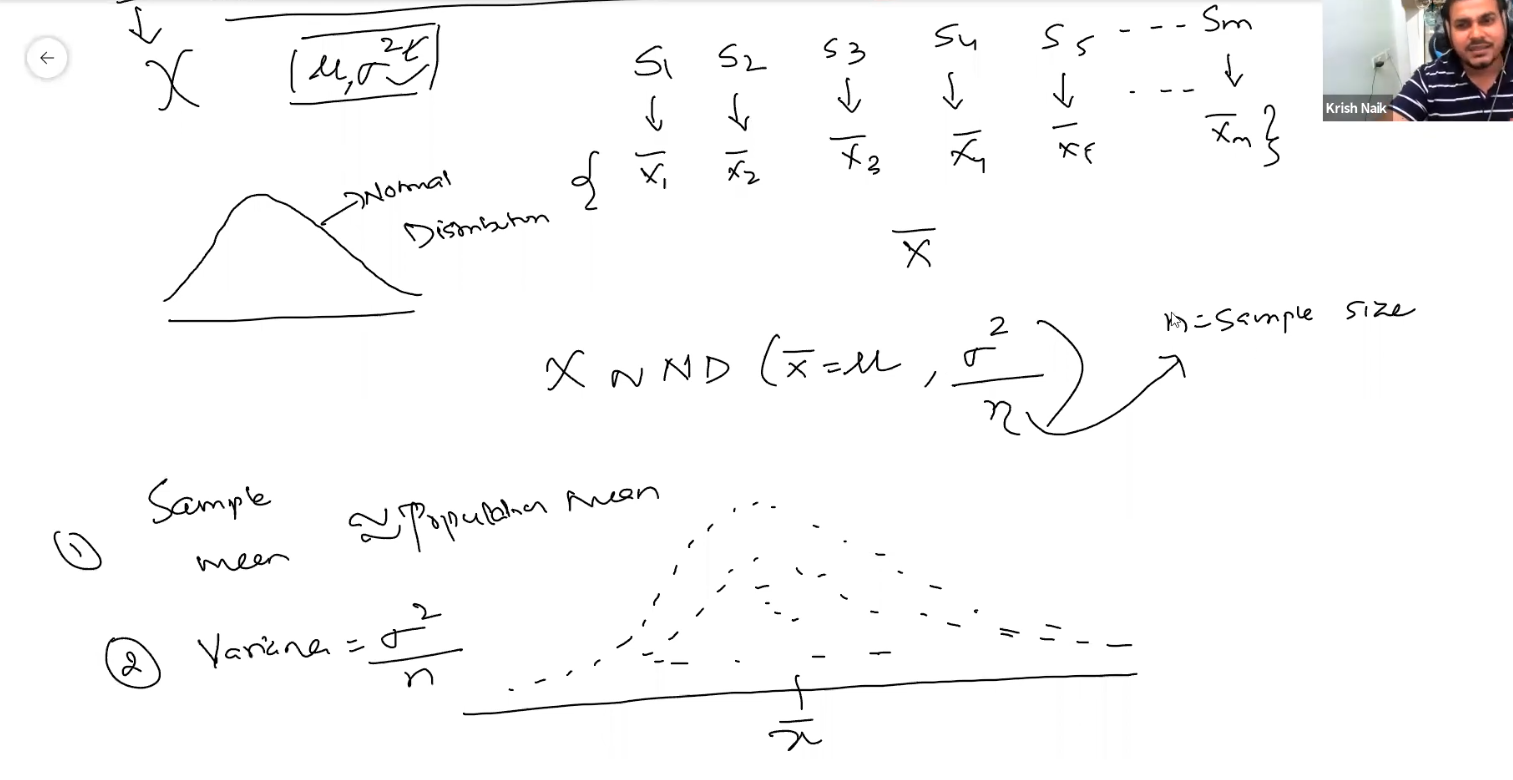
In x the data ranges from 1 to 5, but in y the data ranges from 0 to 4. So with y the computations can be done easy when respect to x.

# Central Limit Theorem

From this population m Samples are derived. From the that m samples individual sample mean is taken out.  
All the sample means 🡪 X

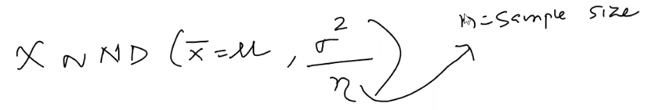


Each sample will have 30 entries



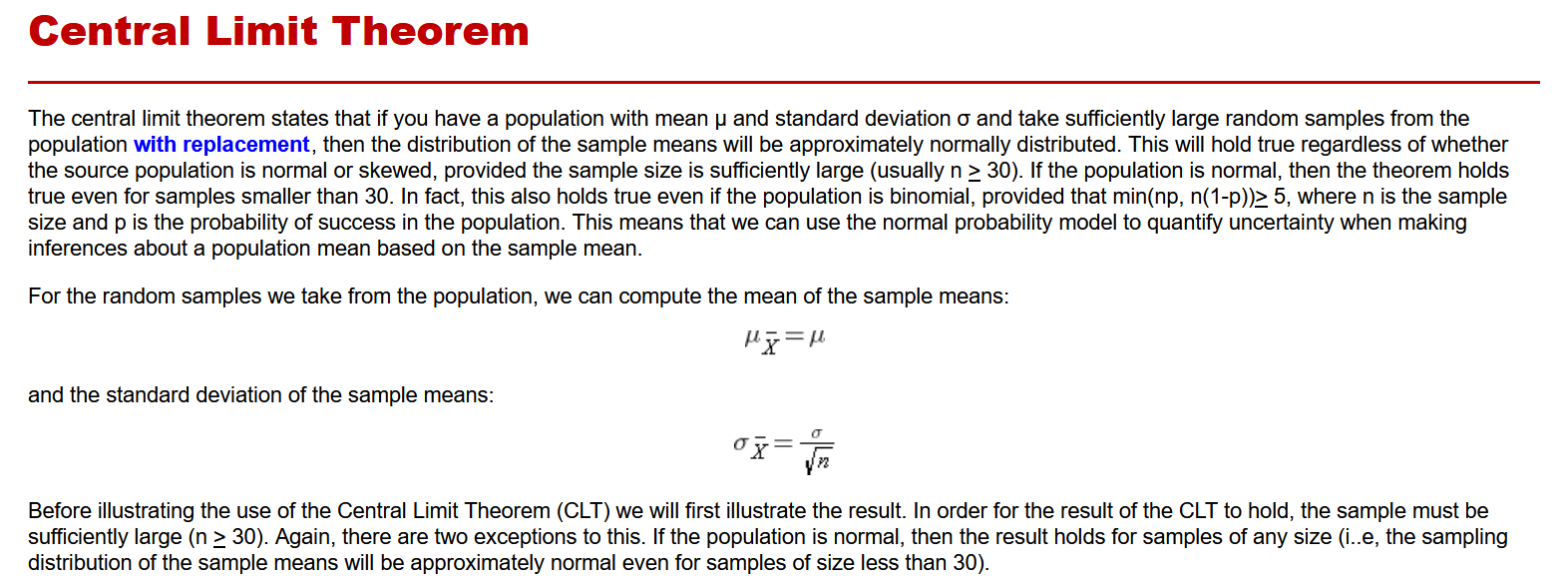
**Why the sample size=30??**

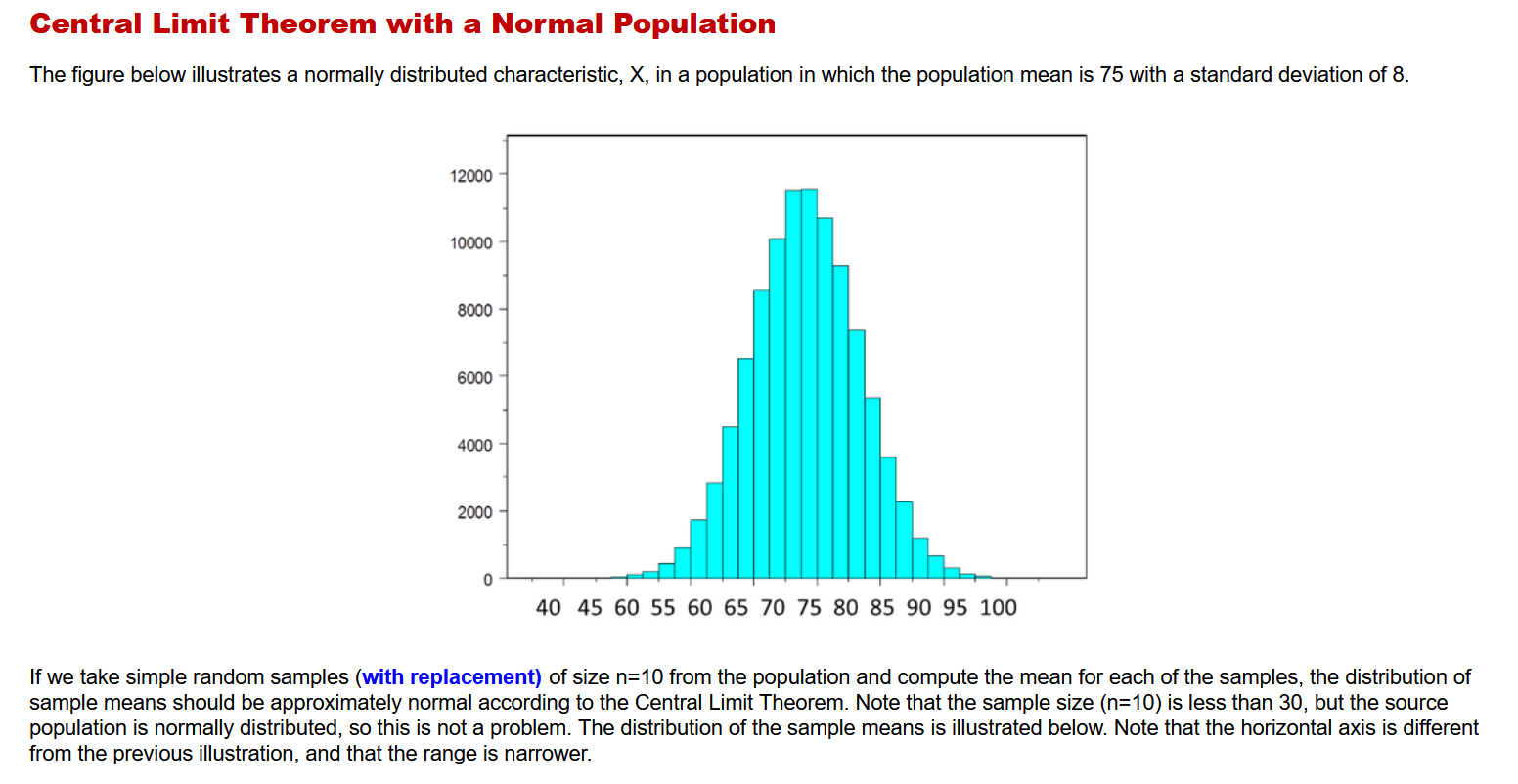
Though Population data-set which doesn’t follow Normal distribution, through various experiments it was found that if we take the sample size for at-least 30(>=30) and we have a also with some mean and variance value, this will follow Normal distribution with mean as µ and variance as (

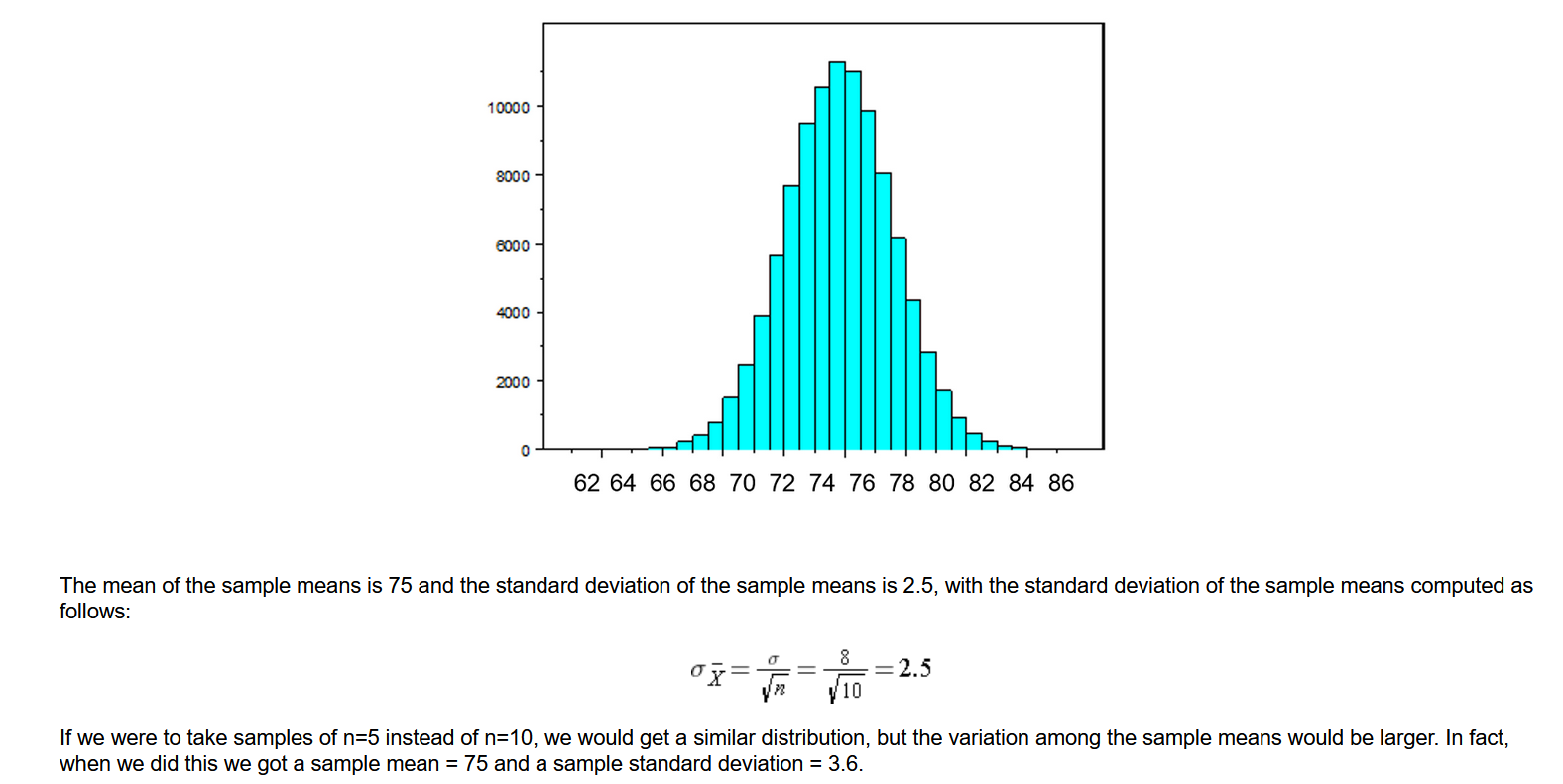


Central Limit theorem states that if   
The Sample distribution of sample mean follows Normal distribution with mean as µ and variance as ( even-though the population follows Normal distribution/Not Normal distribution.  
With this we can apply some techniques with the population to extract useful info.

Below 🡪 <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability12.html>

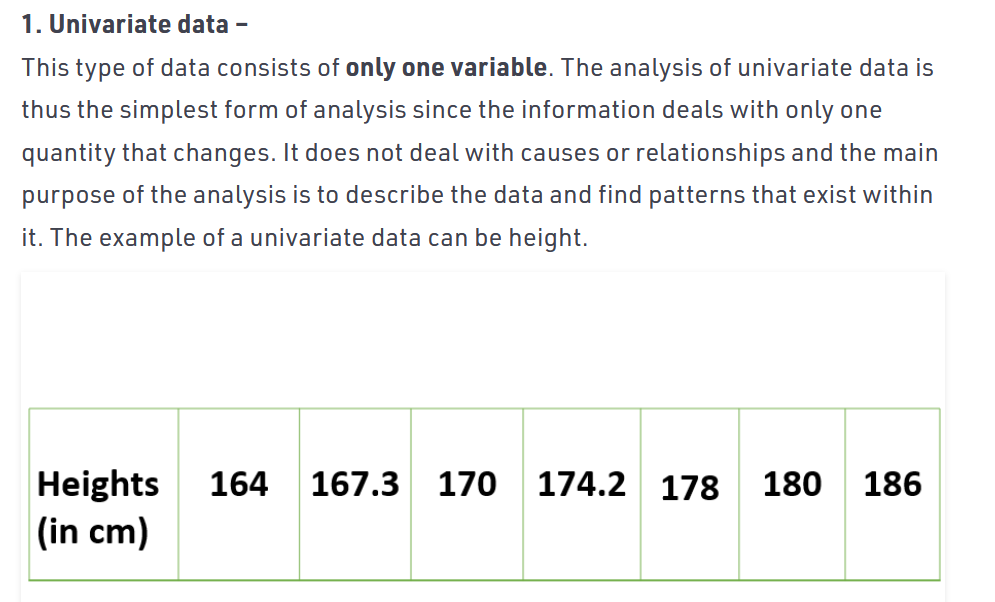


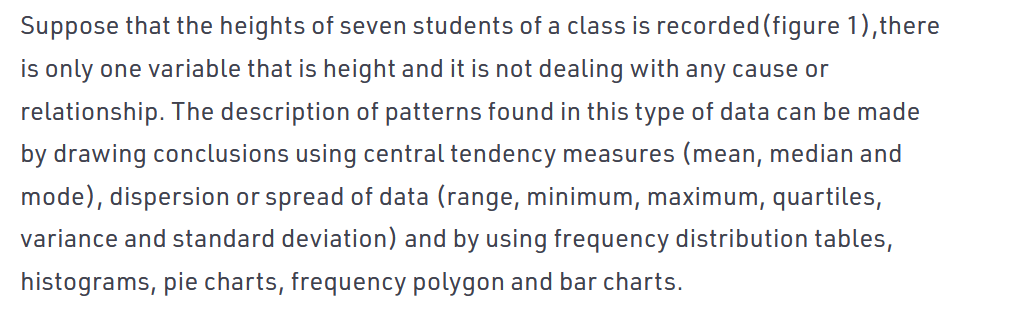




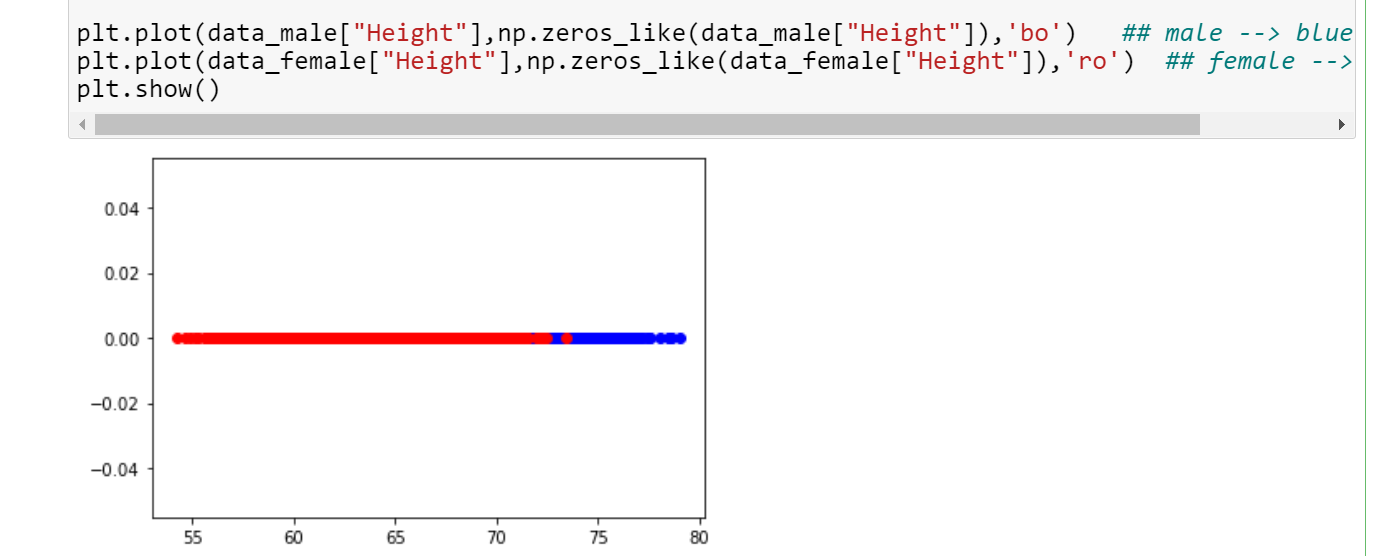
# Univariate, Bivariate and Multi-variate Analysis

## Univariate Analysis

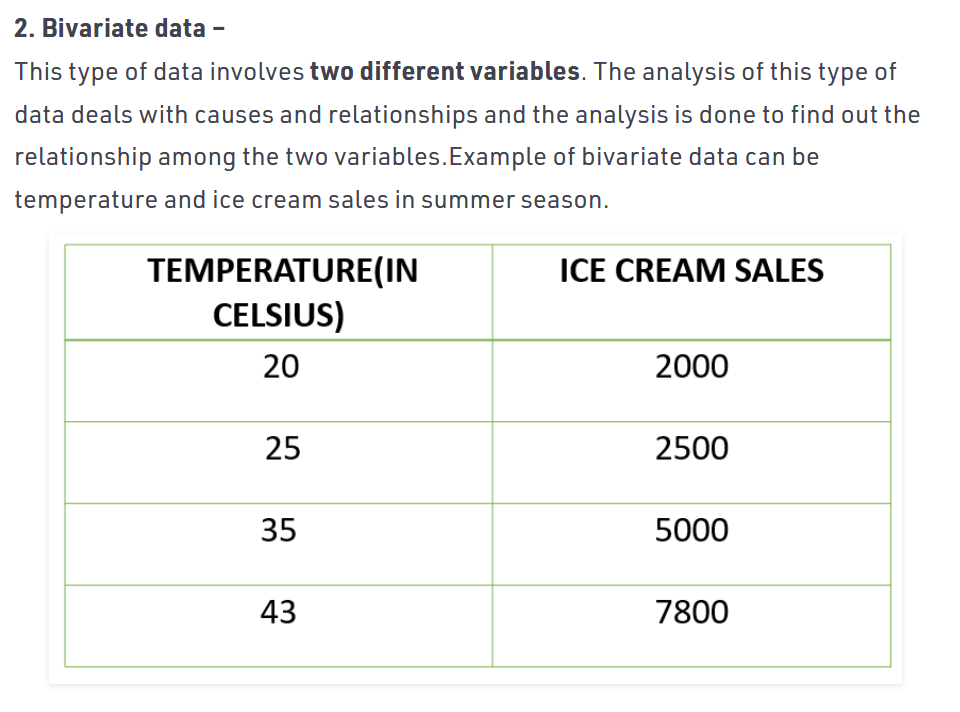


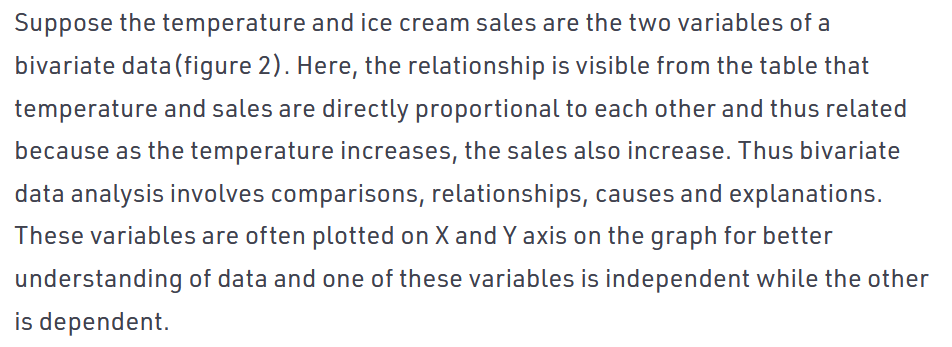


## In Uni-variate there is NO y-axis. Only one dimension, So only we are making Y-axis as 0

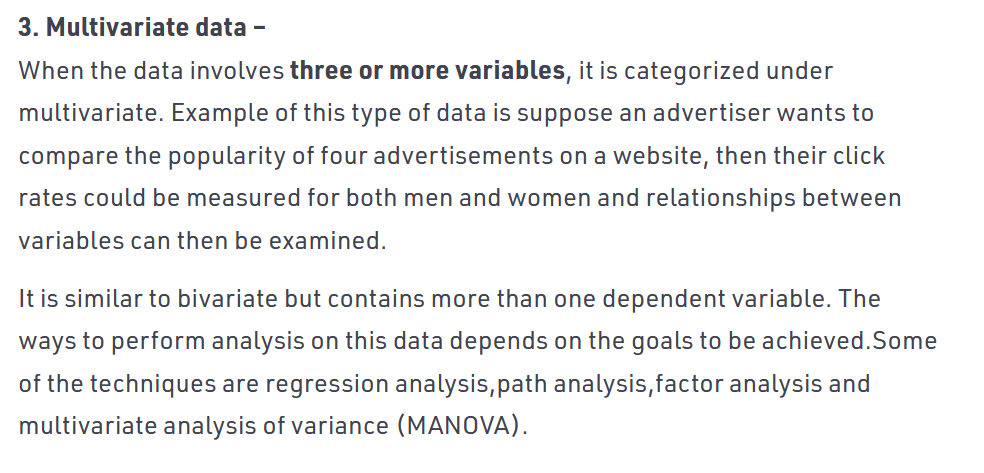


## Bivariate Analysis



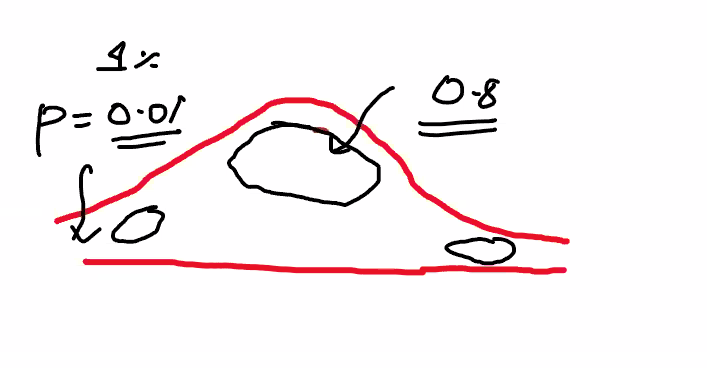


## Multi-variate Analysis



Hypothesis testing

It will be used in EDA purpose.



p 🡪 Probability of pressing that particular area

p=0.01 (1% of touching the key) These are taken from sample  
p=0.8 (80% of touching the key)

p value is the probability of NULL hypothesis to be TRUE.

**One-tailed and two tailed test**

**Statistical and Practical significance**

**Independent samples and Dependent samples**