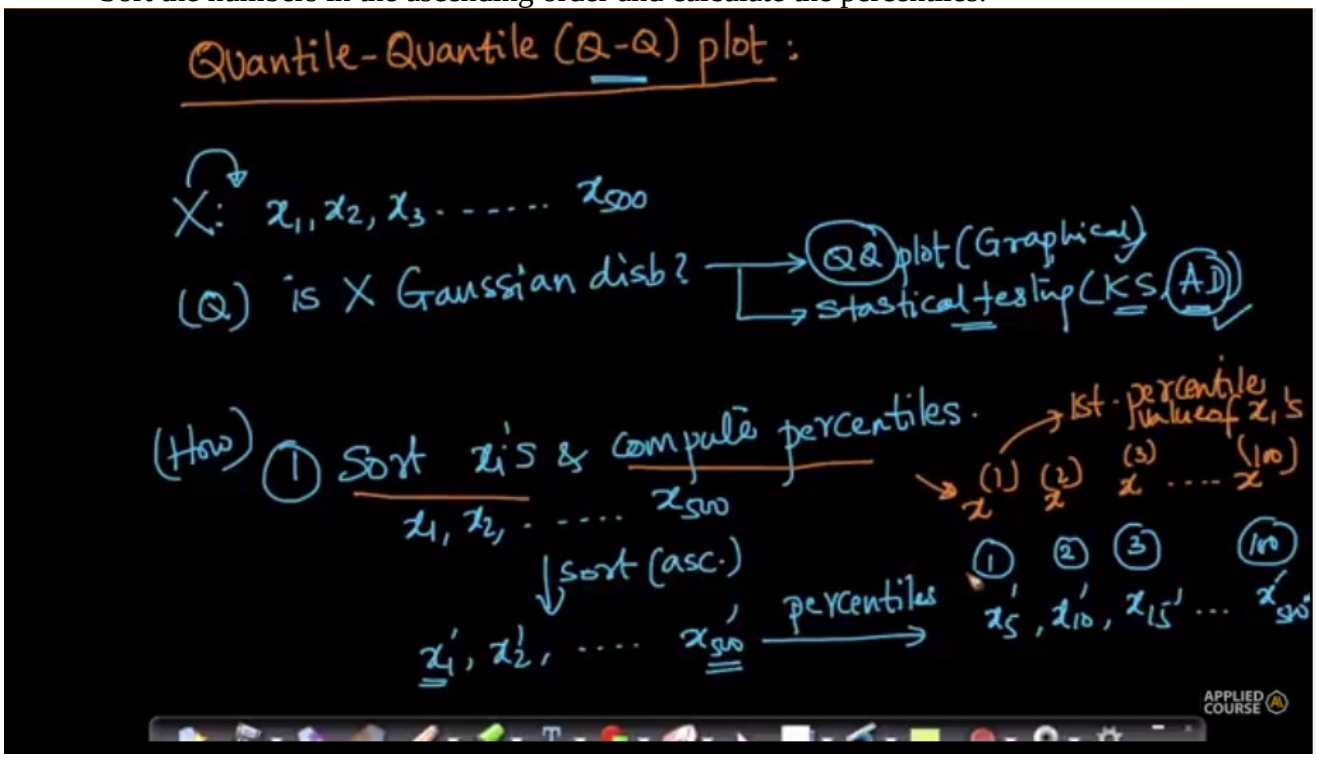
NOTES

Q-Q plot:

Step – 1

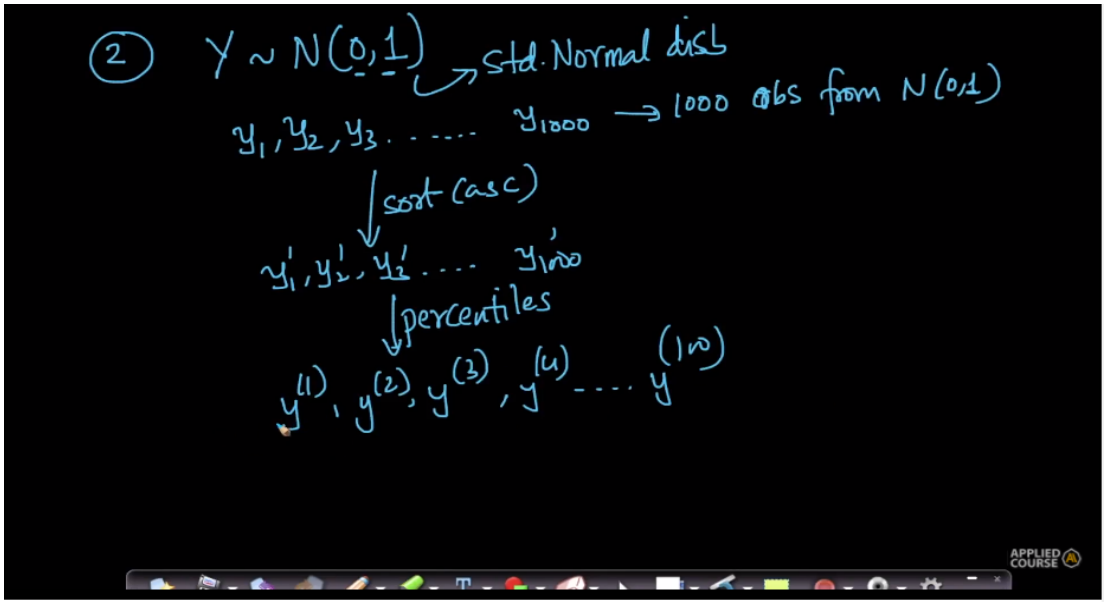
The x1 to x500 are the random variables of the data points. The goal is to check whether the data is normally distributed.

- Sort the numbers in the ascending order and calculate the percentiles.



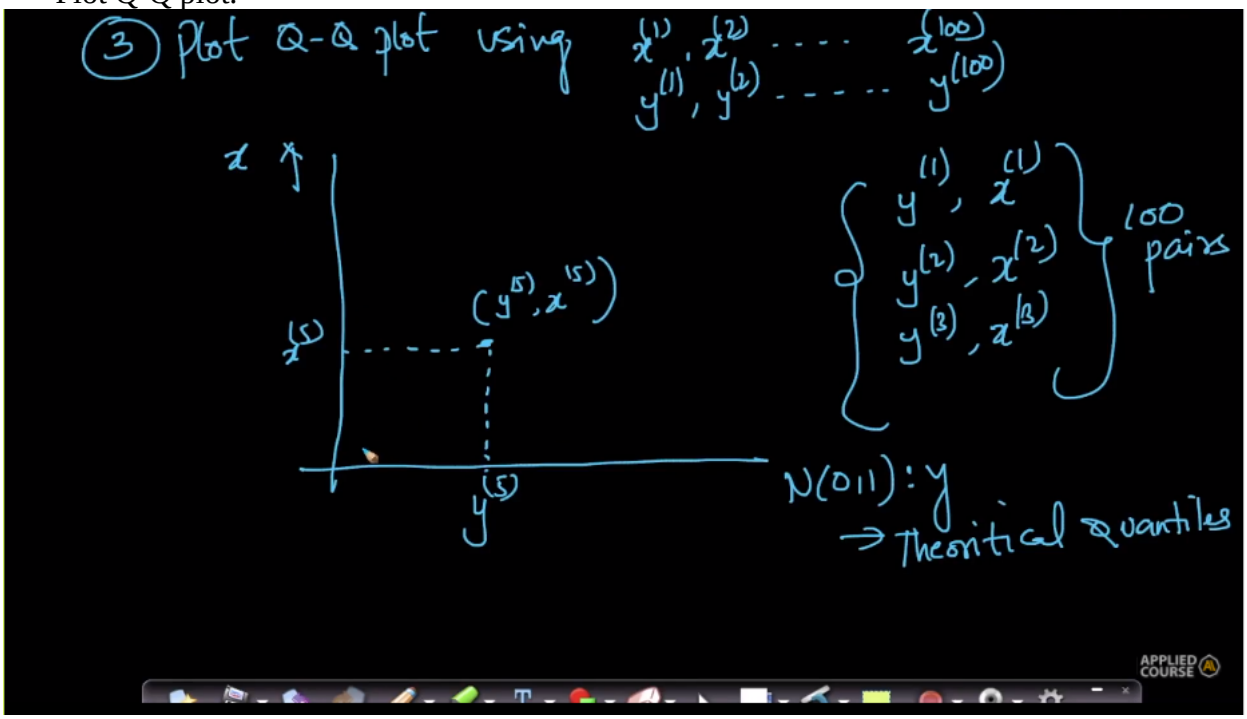
Step – 2

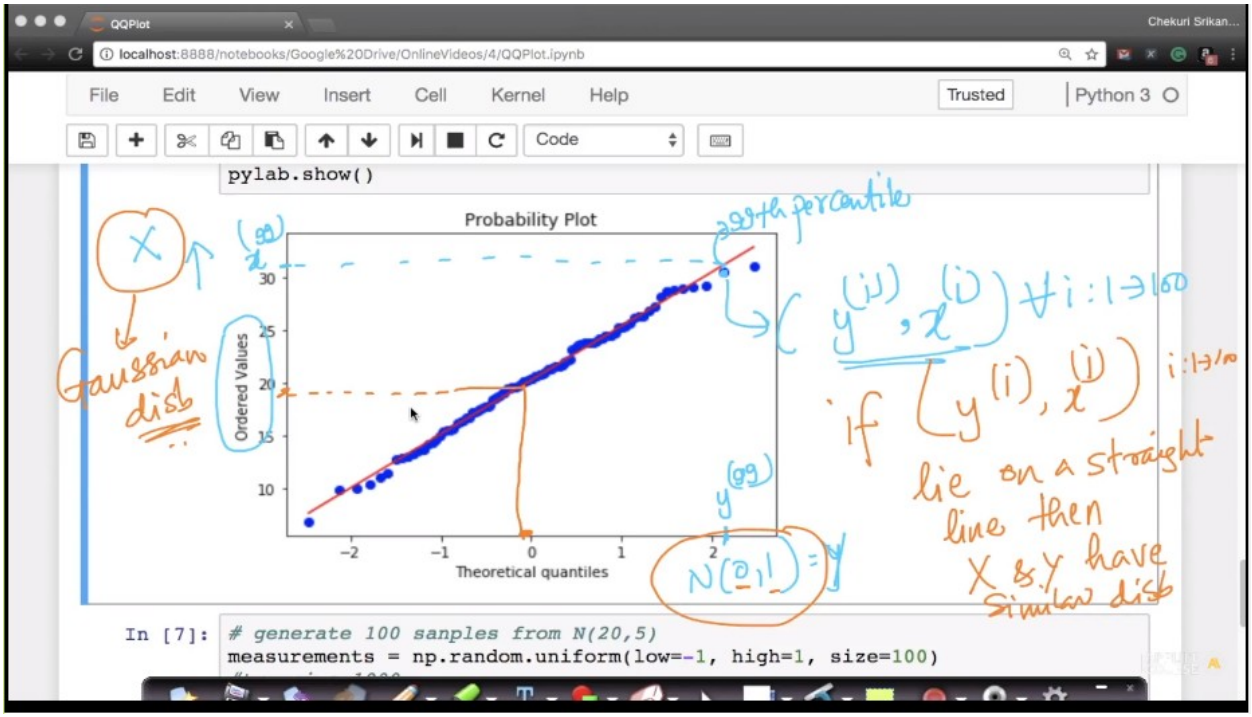
- Take the random observations from the standard normal distribution and repeat the process in the step – 1.



Step – 3

- Plot Q-Q plot.





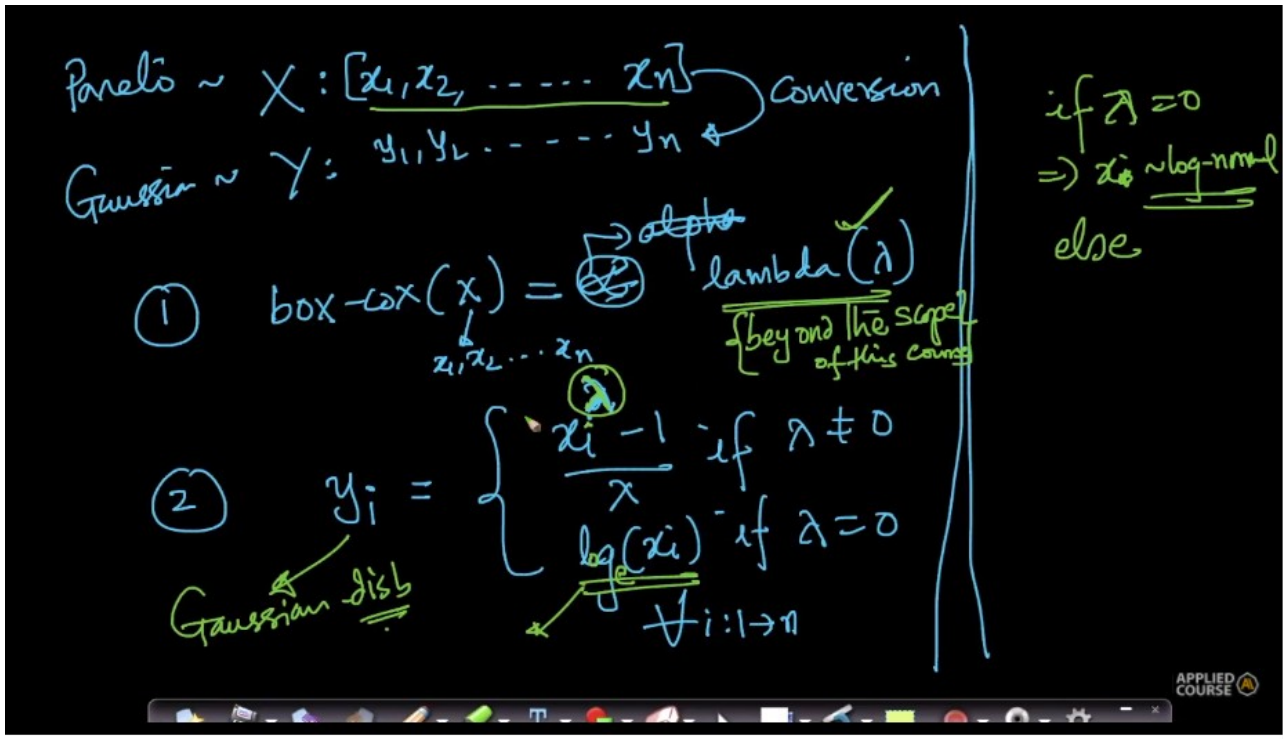
- If the points lie on the same line then the data is from the same distribution, When ever the samples or observations is very hard to interpret Q-Q plot.

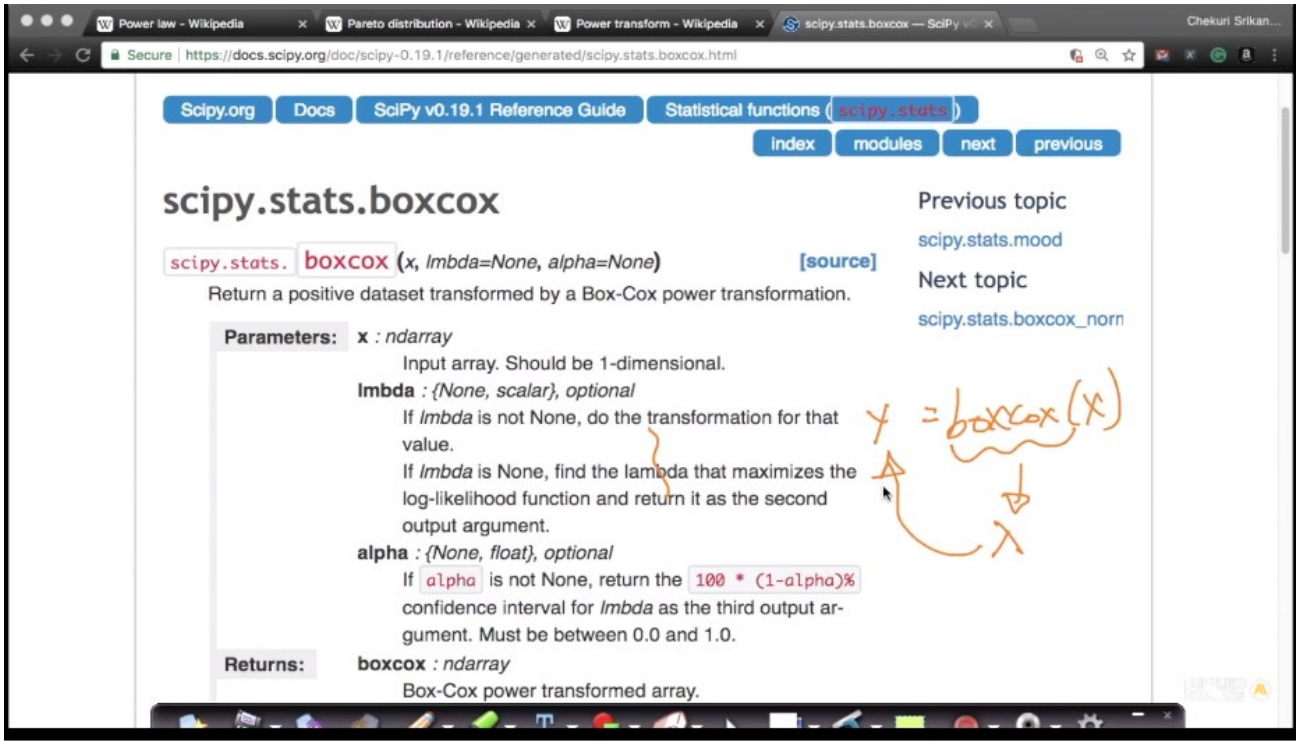
Power – law (Box – Cox transform)

- **Log-normal** distribution is converted to gaussian by taking natural logarithm of the random variable or data points in **log-noraml** distribution.

Power – transform(conversion of power law distribution to Gaussian distribution)

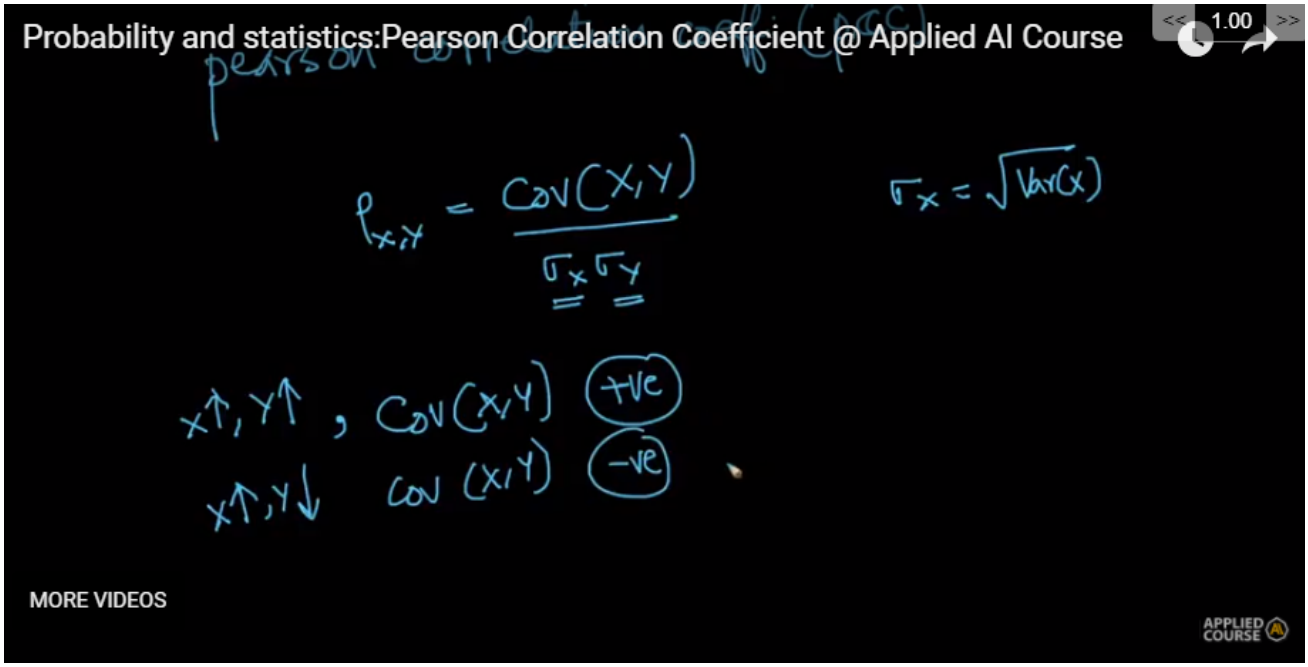
- Pareto to Gaussian.





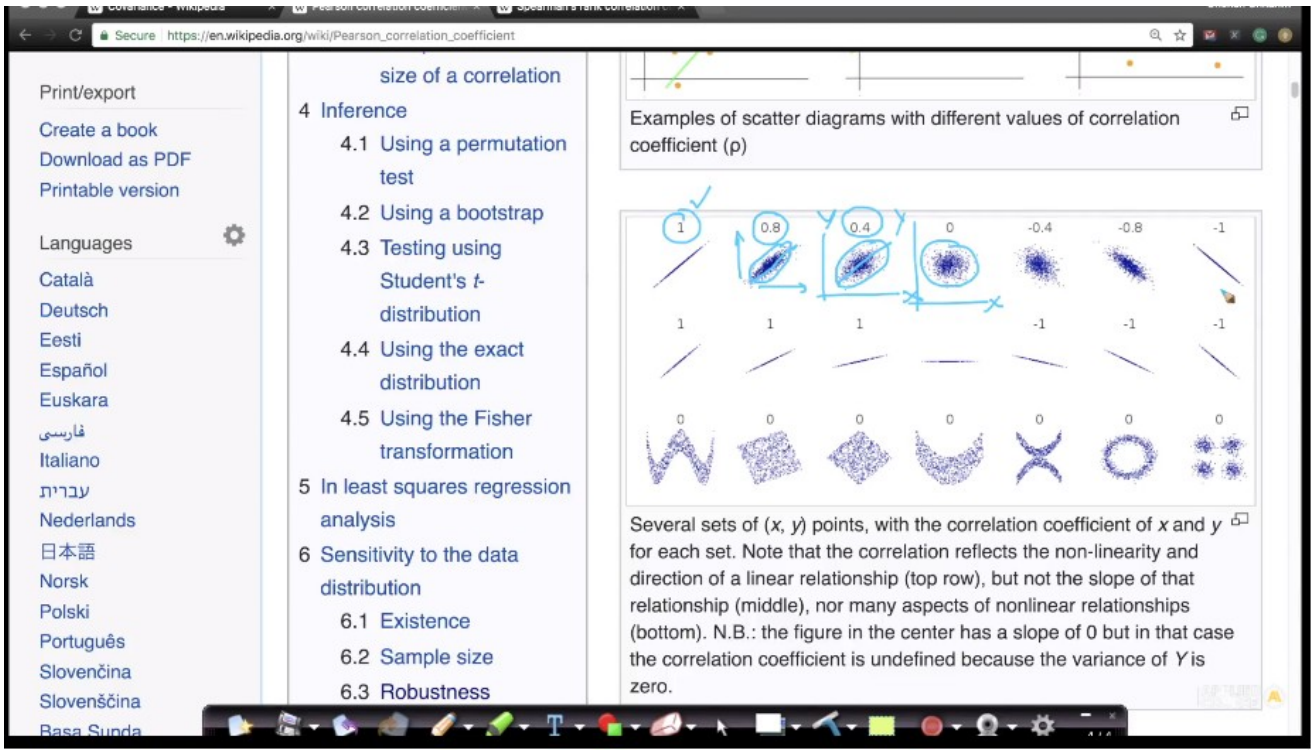
Pearsons correlation coefficient:

- This lies in the range of -1 and +1.

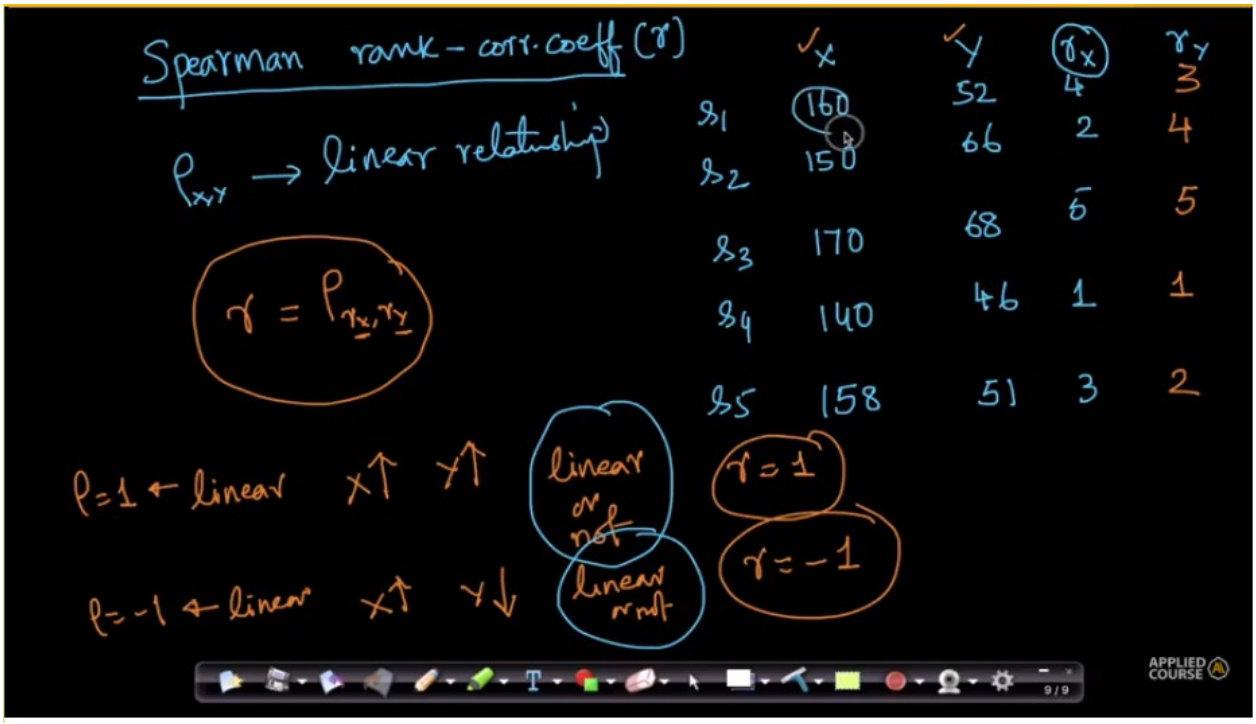


Limitations:

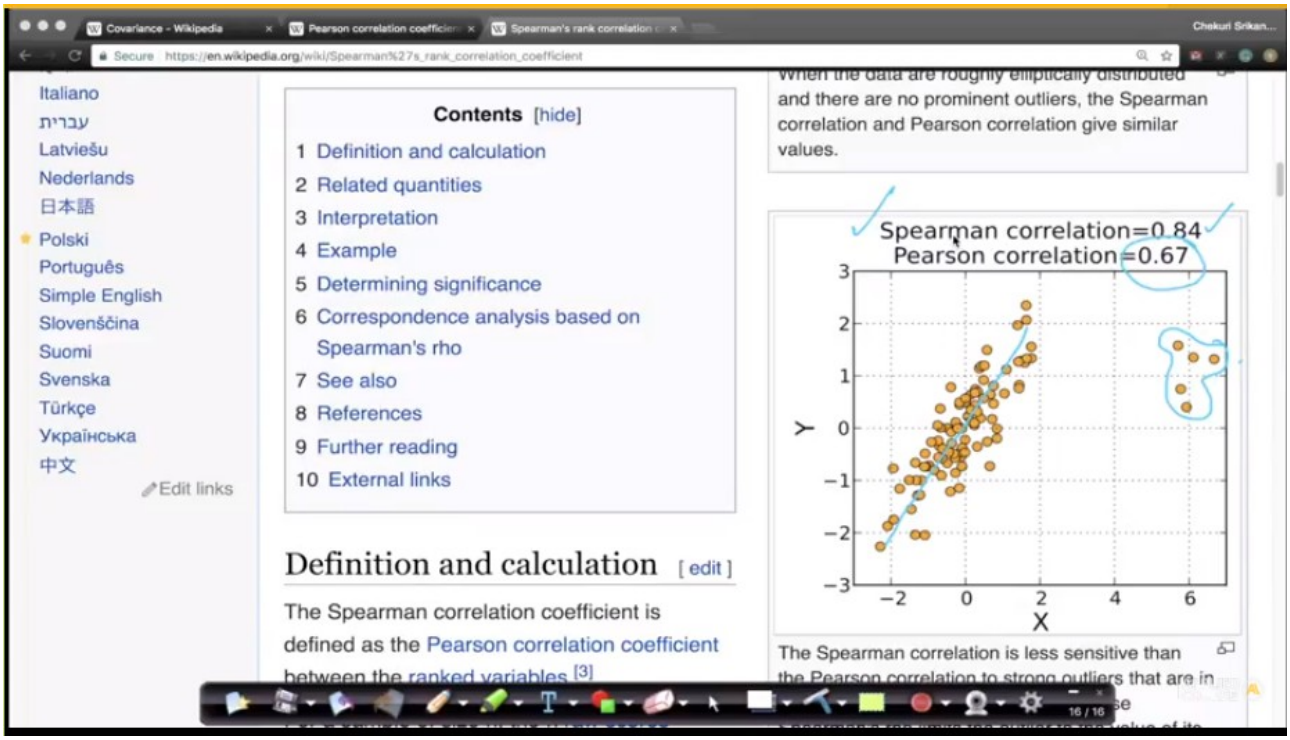
pearson's correlation coefficient does not give any interpretation for the data if it is in shapes.



Spearman's rank – correlation coefficient:



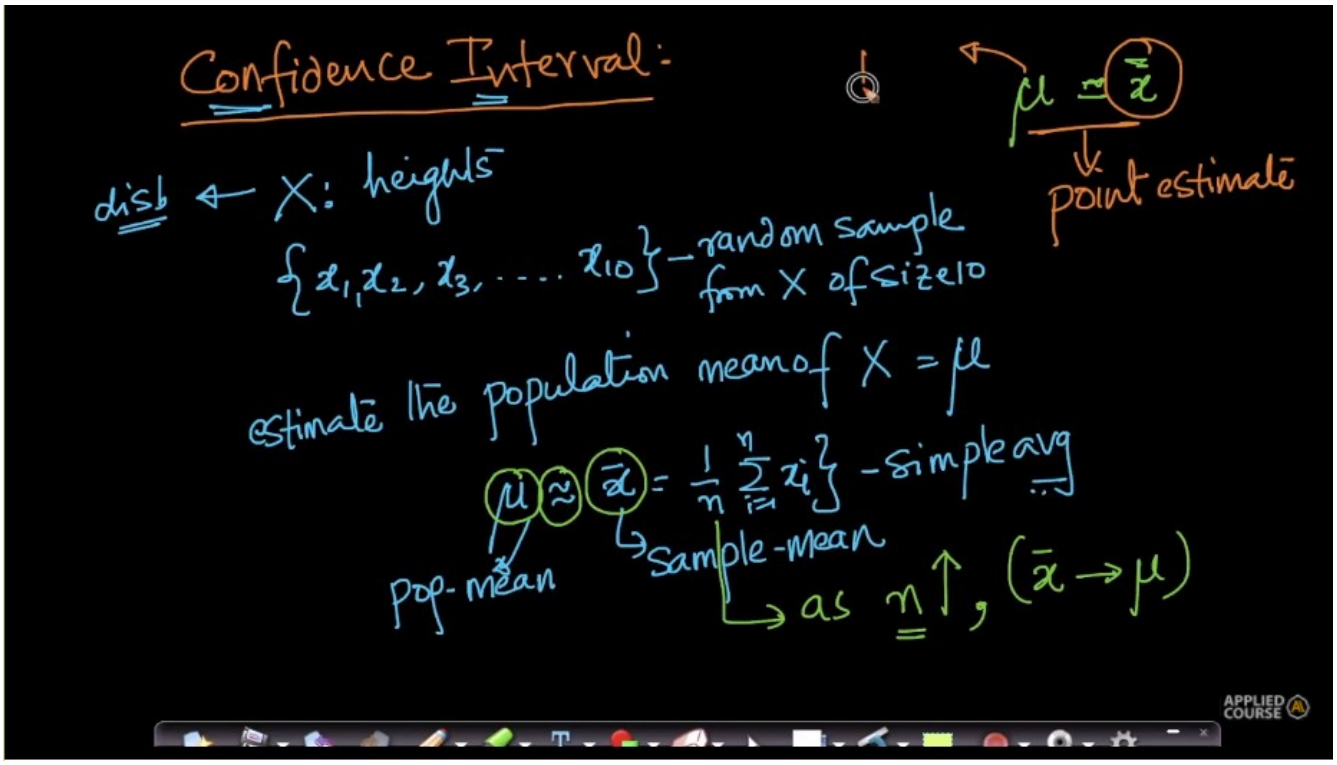
Spearman's correlation coefficient does not take into consideration the linear or not.



Confidence intervals:

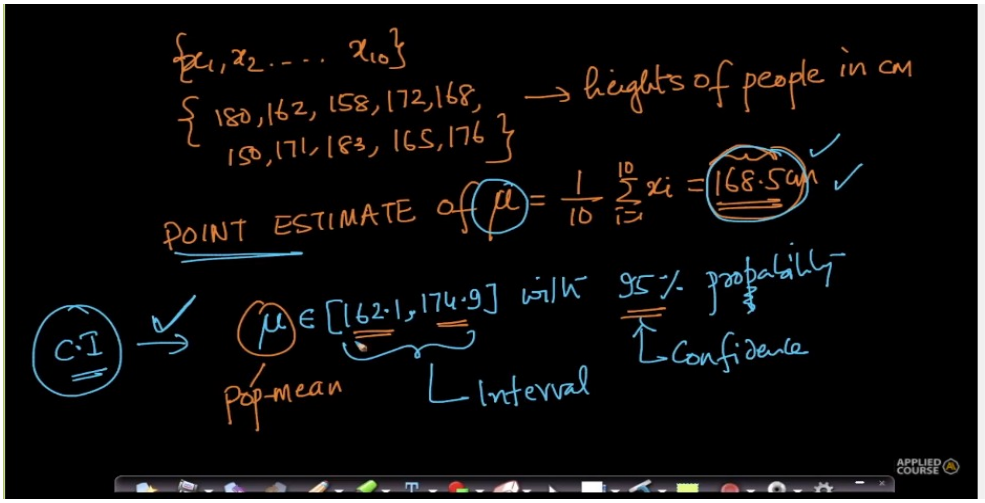
point estimate:

Estimating the sample mean to a single value by using the population mean this is called point estimate.



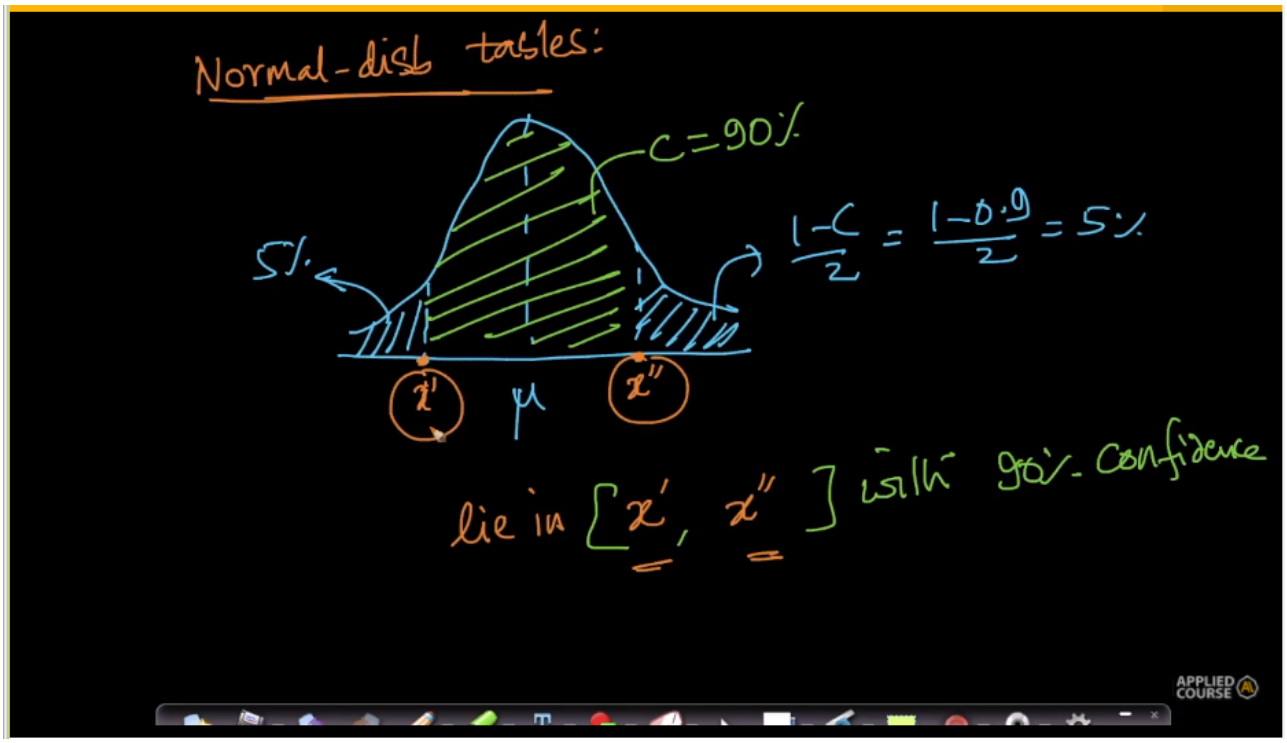
There is a better way of interpreting the sample mean using the **confidence intervals.**

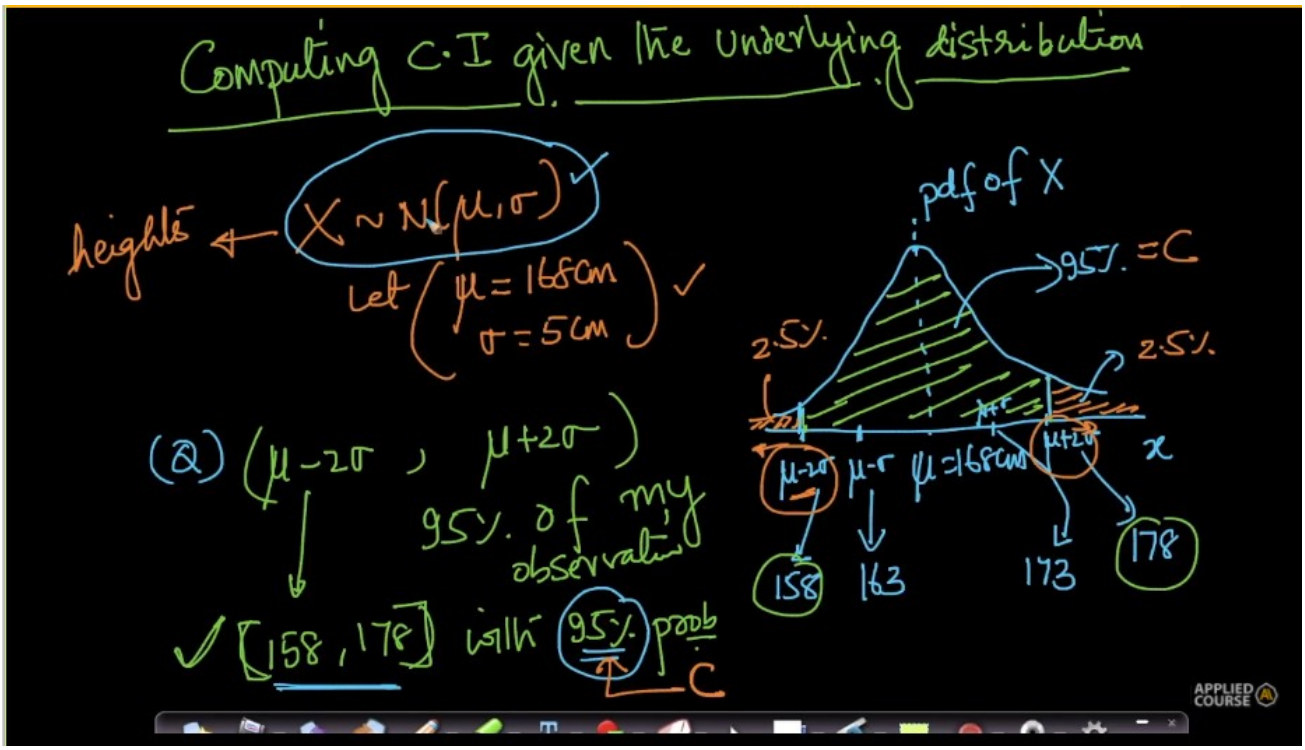
Confidence intervals:
Confidence intervals gives the confidence of the mean value associated with the probability score.
Intro:



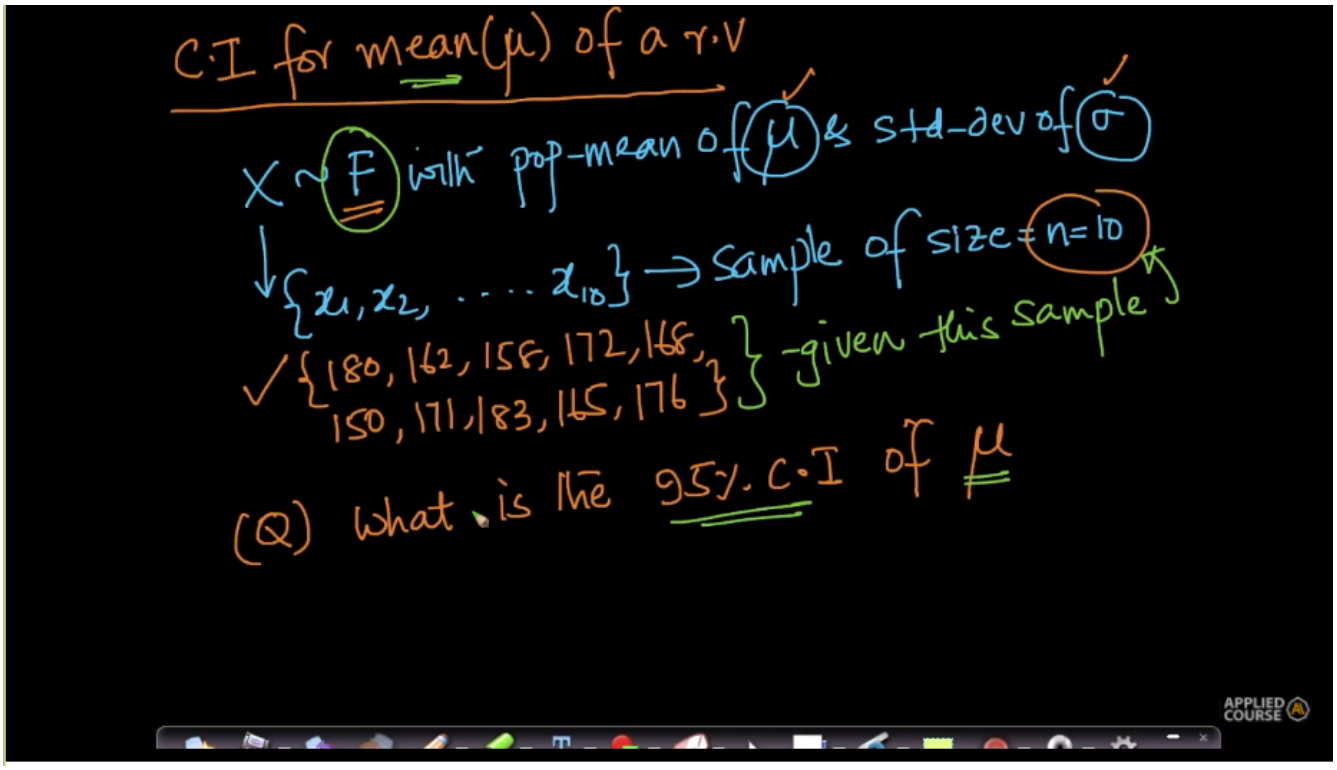
Calculation of confidence intervals:

Intro:





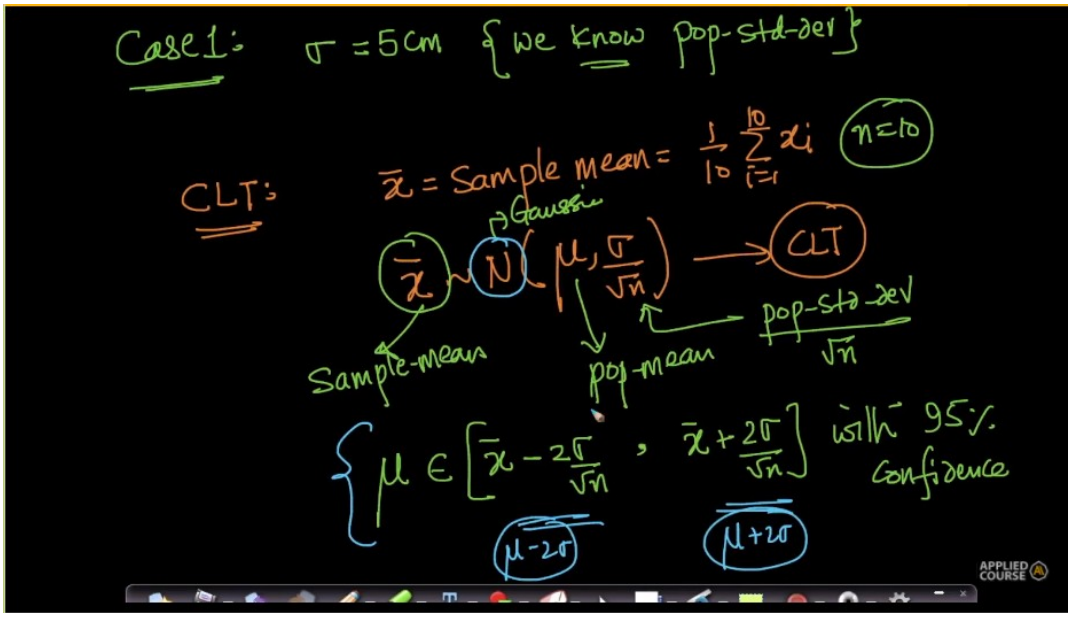
calculation:



Central limit theorm:

- Central limit theorem says that the **mean** and **standard deviation** of the sample is given as follows:

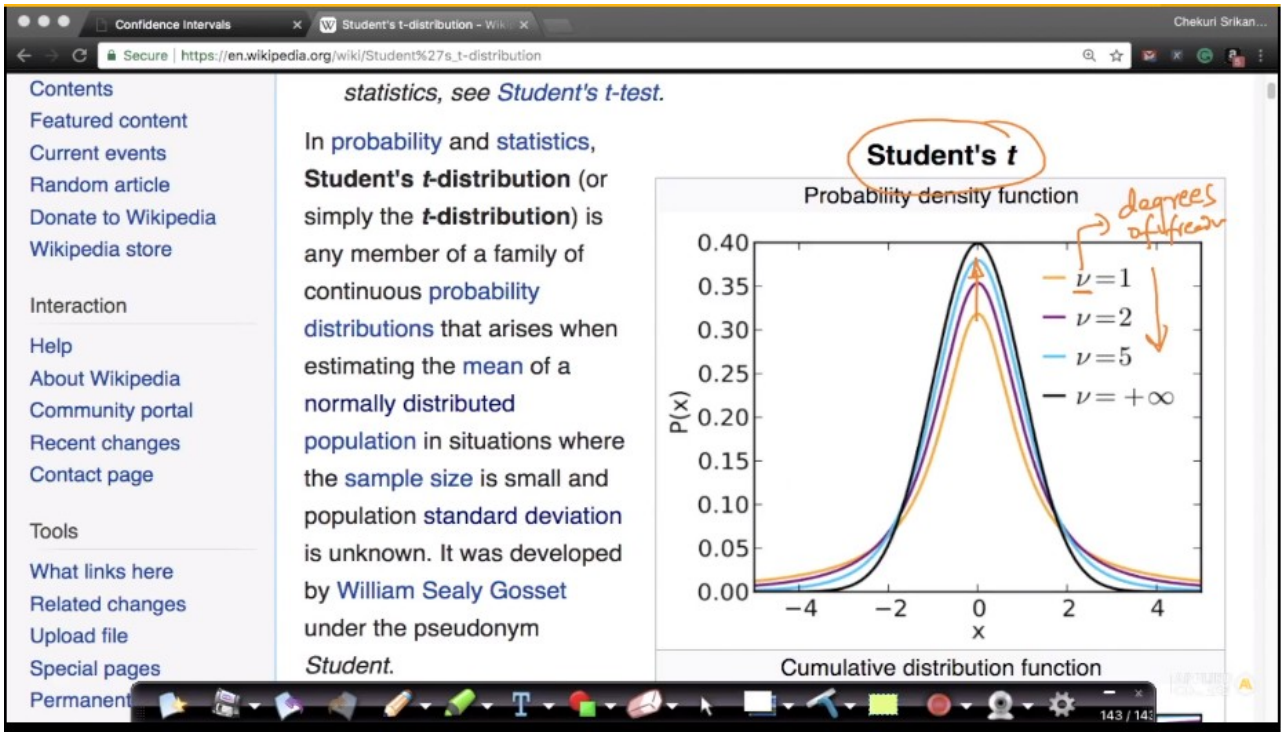
Case – 1(you have <sigma> and <mean> of population):



- The sample can come from any distribution as long as the population has finite **mean** and **variance.**

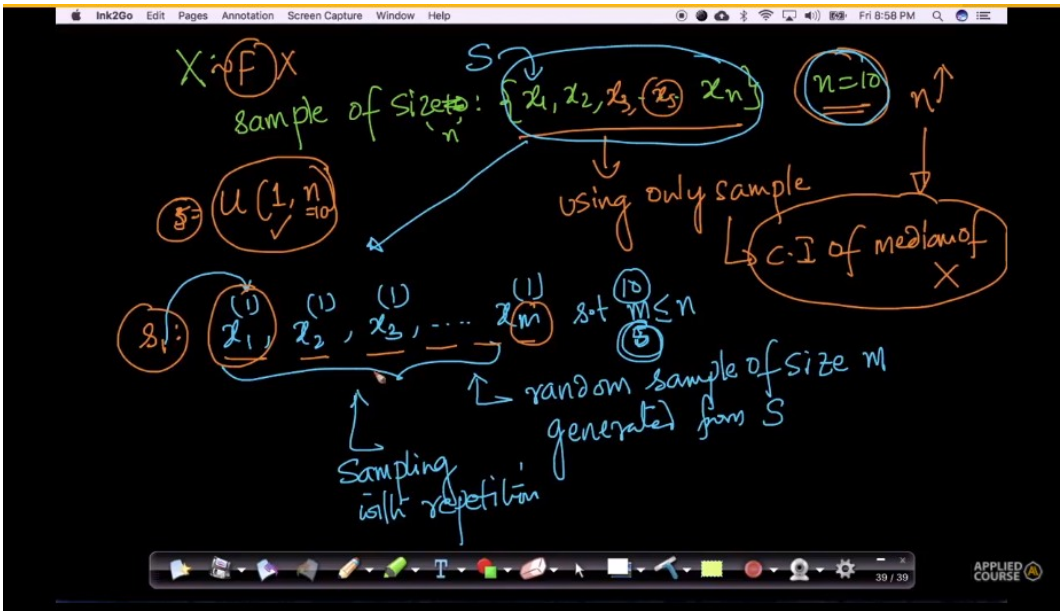
Case – 2:(If we do not know population standard deviation)

We use students t – distribution with n-1 degrees of freedom where n is the size of the sample.

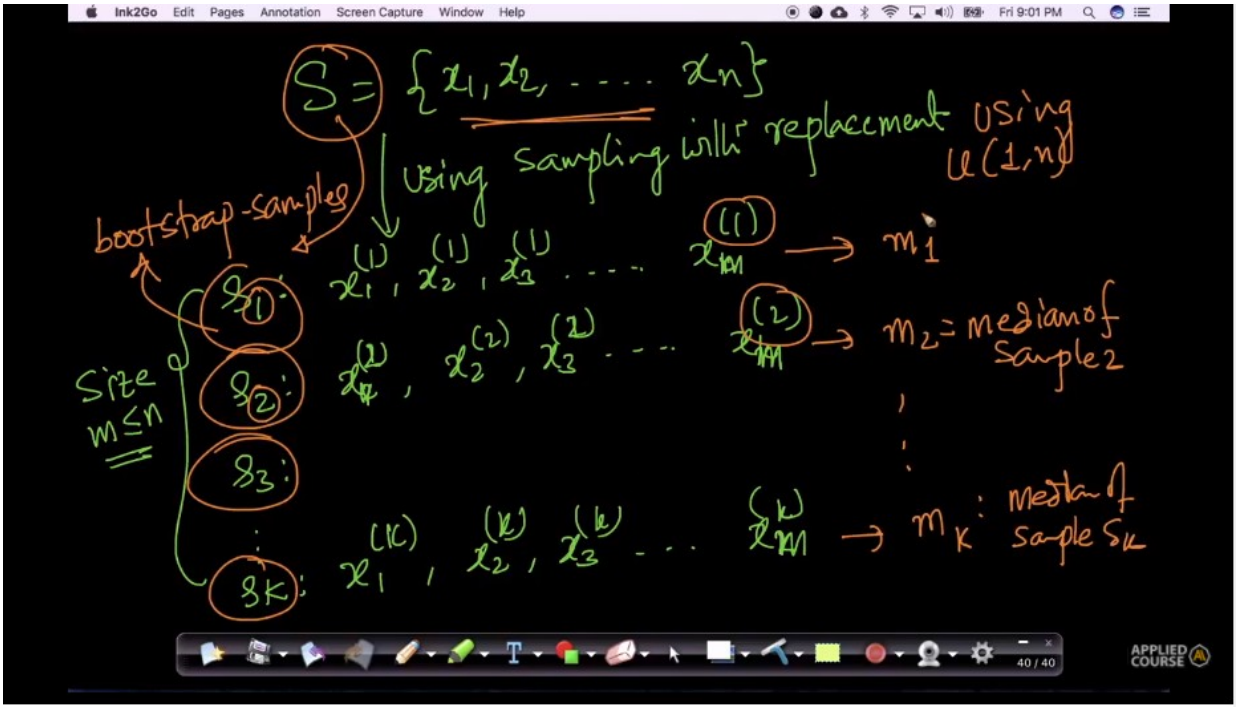


Confidence interval using bootstrapping:

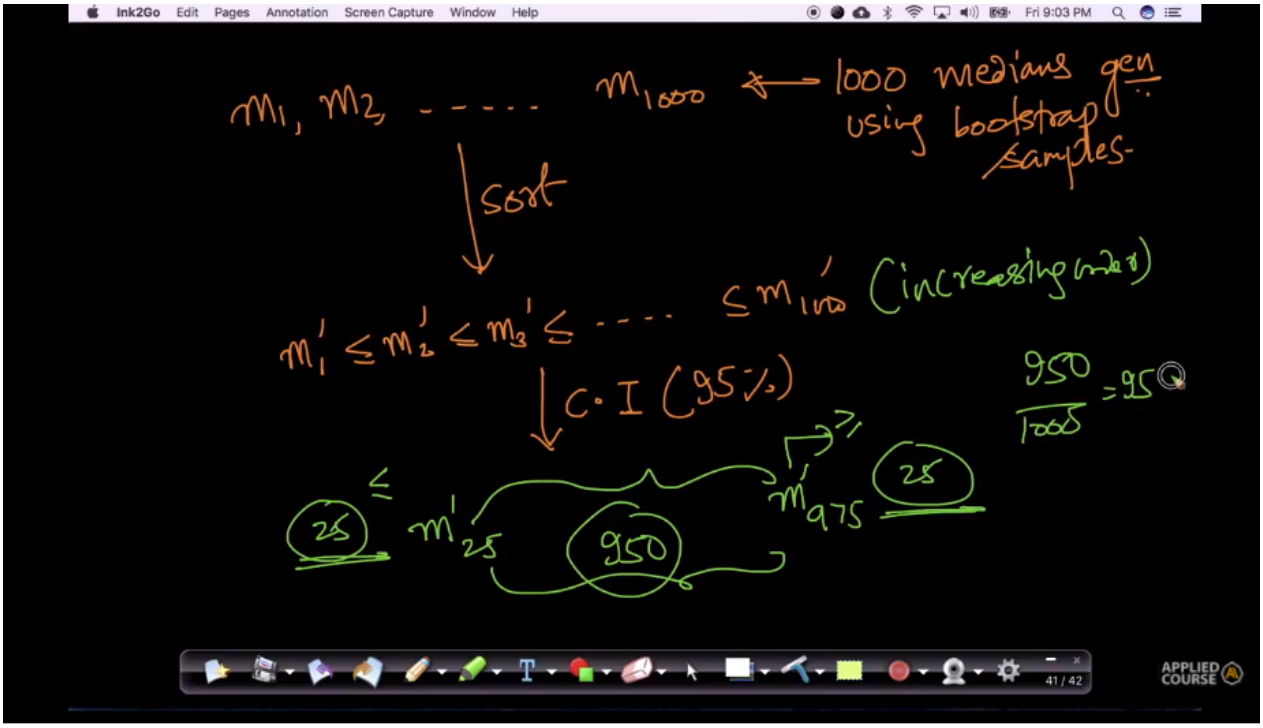
- The sampling is done by replacement, for choosing the values we can use uniform distribution.



- We make many samples using the same methodology to make many samples these are called bootstrap samples.



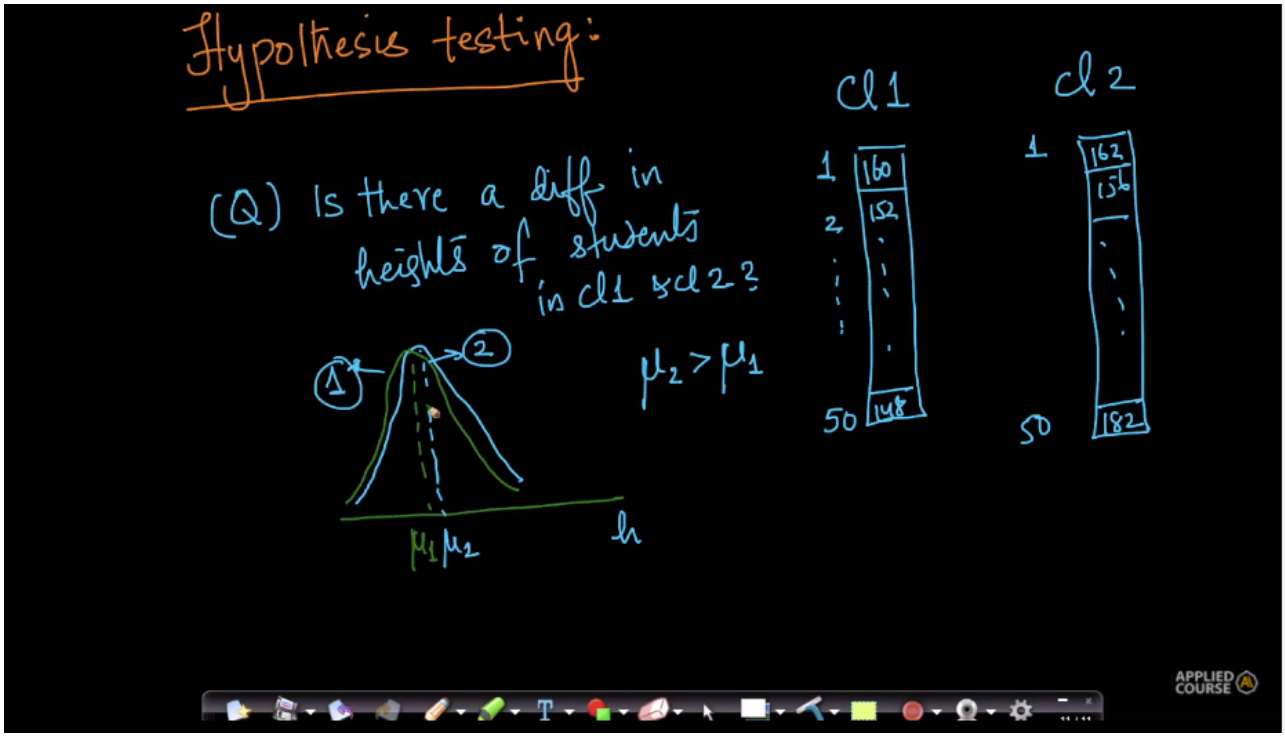
- For calculation of median of the sample, first the medians of the bootstrap samples are made.
- Then the medians are sorted in increasing order.
  Then the confidence interval is calculated, here m25 and m975 are the 95 percent CI for median of the sample.



Similar calculation is made for calculating the **variance, mean** and **standard deviation.**

THIS METHOD OF CALCULATION IS CALLED NON – PARAMETRIC TECHNIQUE, DOES NOT MAKE ANY ASSUMPTIONS ABOUT THE DATA.

HYPOTHESIS TESTING:

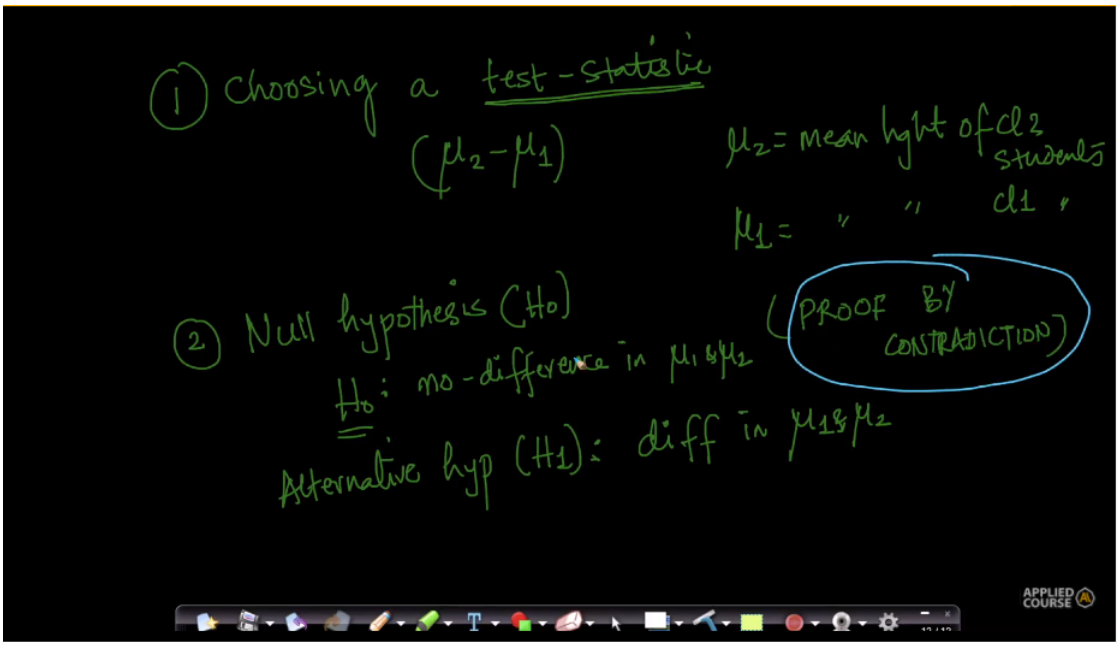


Step – 1

Choosing the test – statistic.

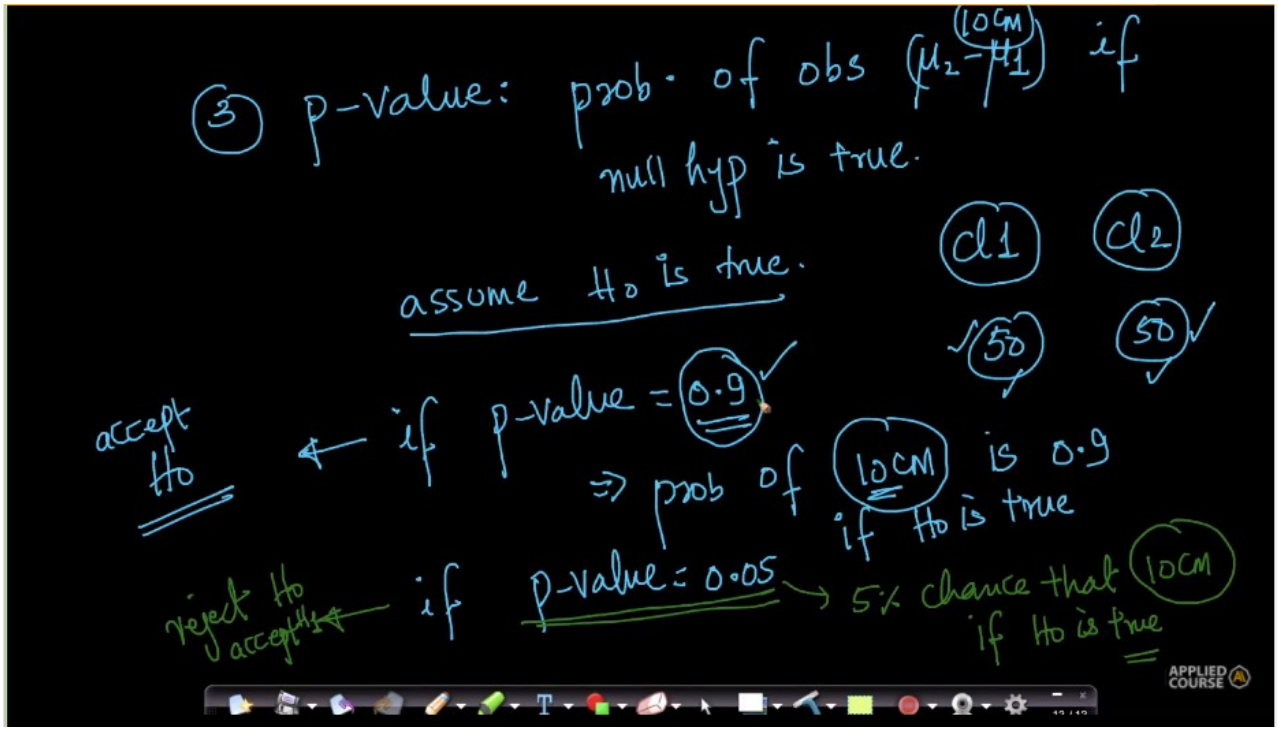
Step – 2

Choosing null – hypothesis.



Step – 3

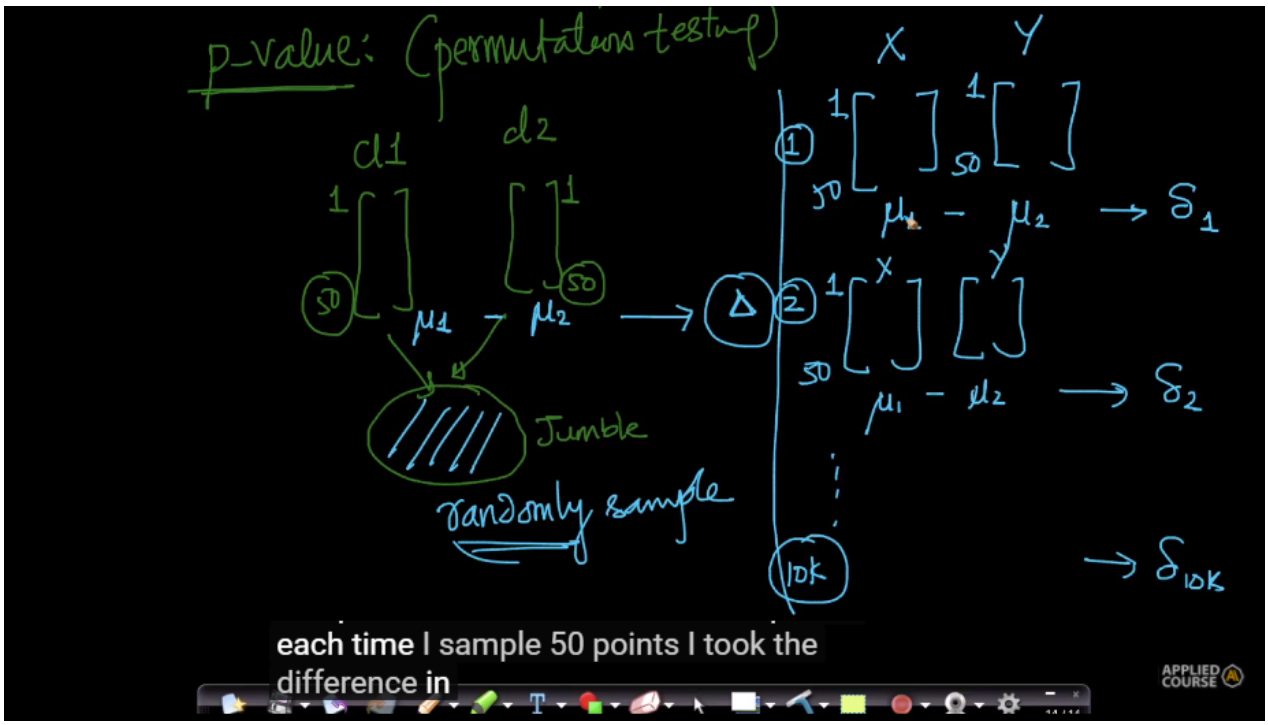
Computing the p – value.



How to compute p – value:

- Step – 1

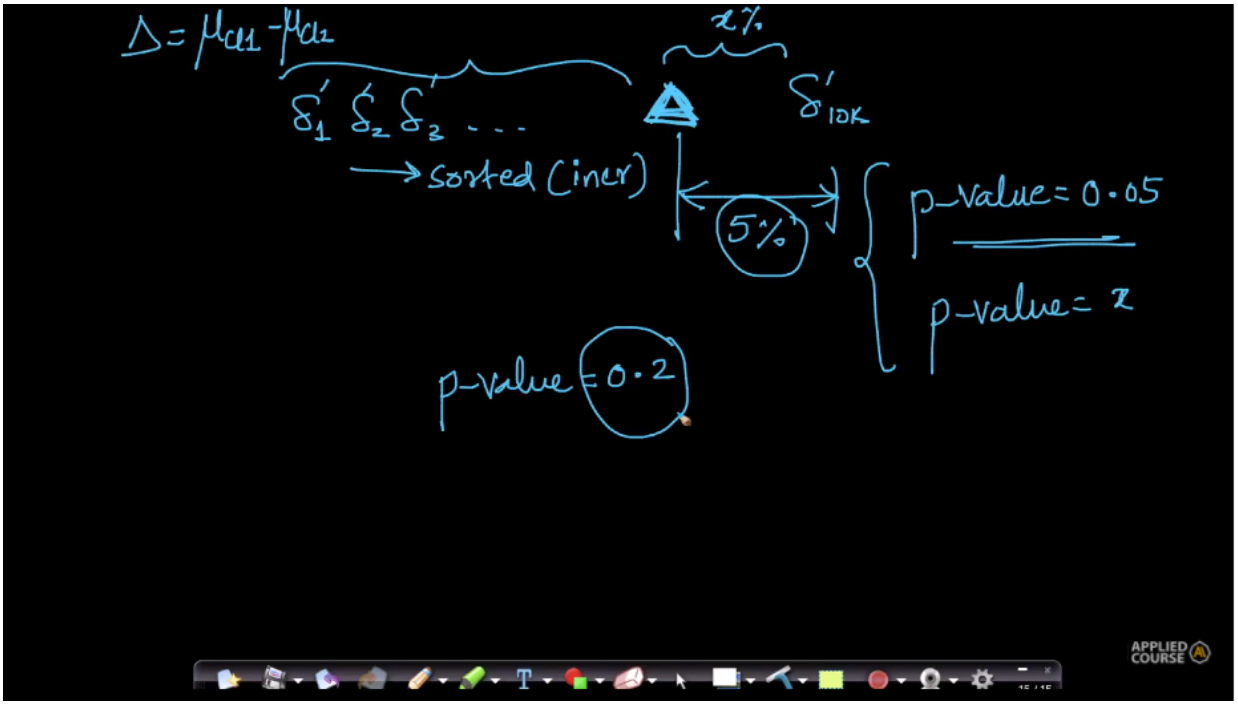
  Randomly make the samples of from the **jumbled data** of the **two** original **classes.**



- Step – 2

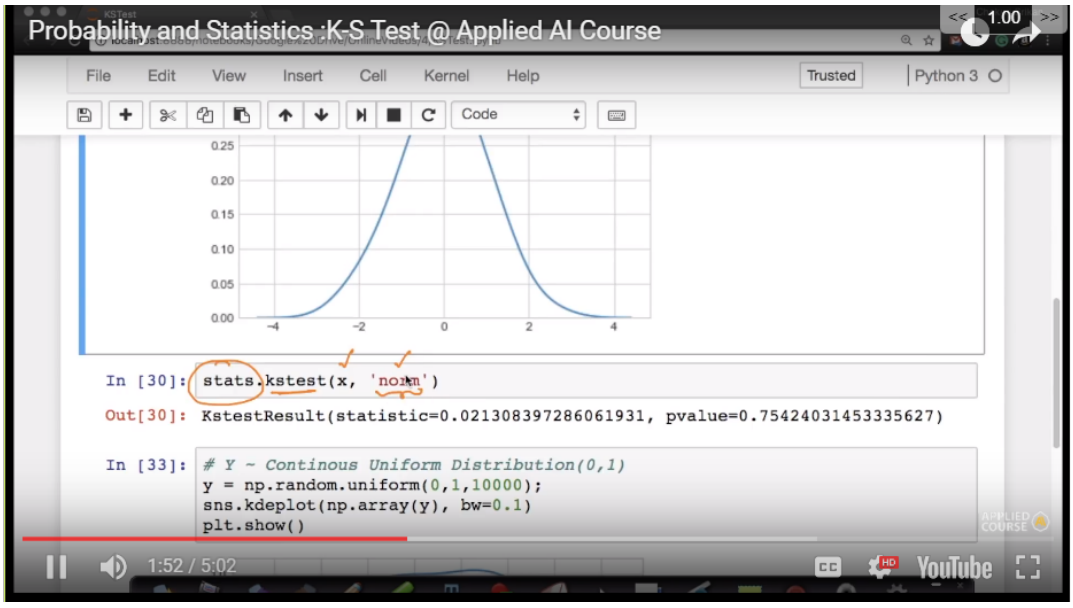
  Then sort the difference of the **mean statistic** that is chosen from the sampled classes. That is all **delta's.**

- Step – 3



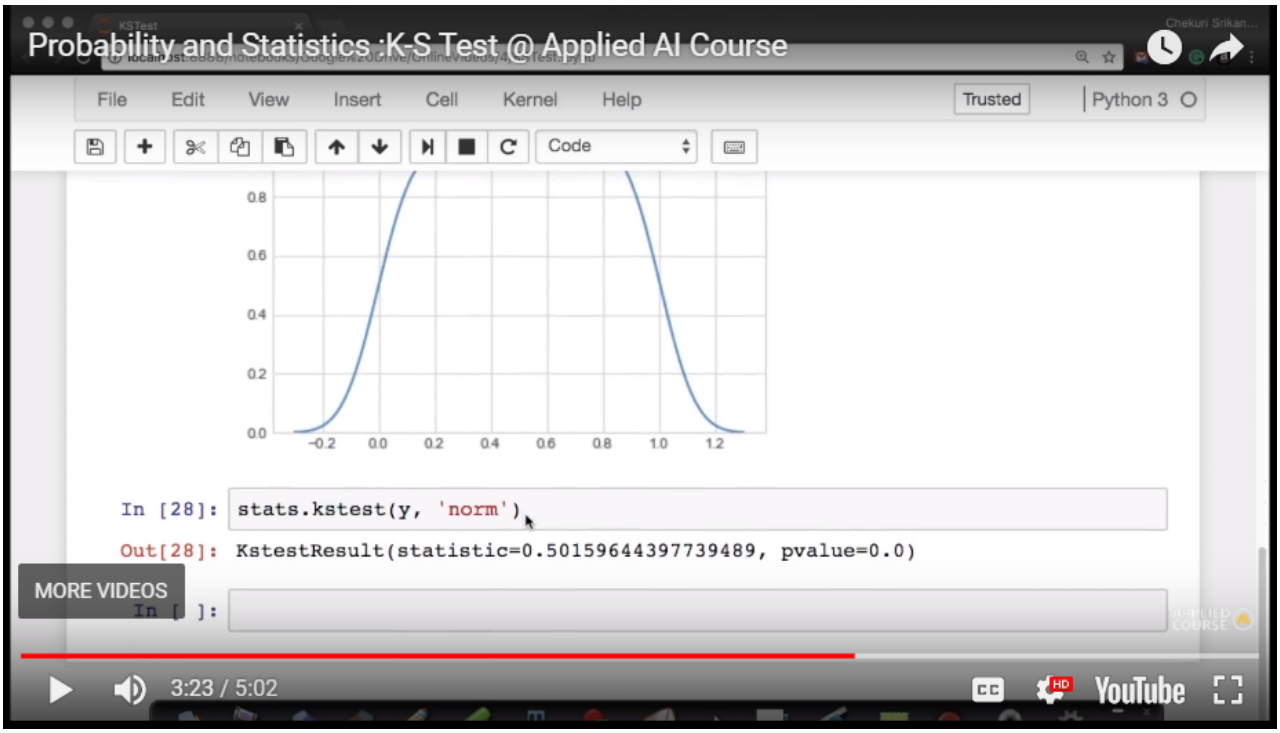
K-S test:

This test is for similarly of two distributions.
For normal distribution.



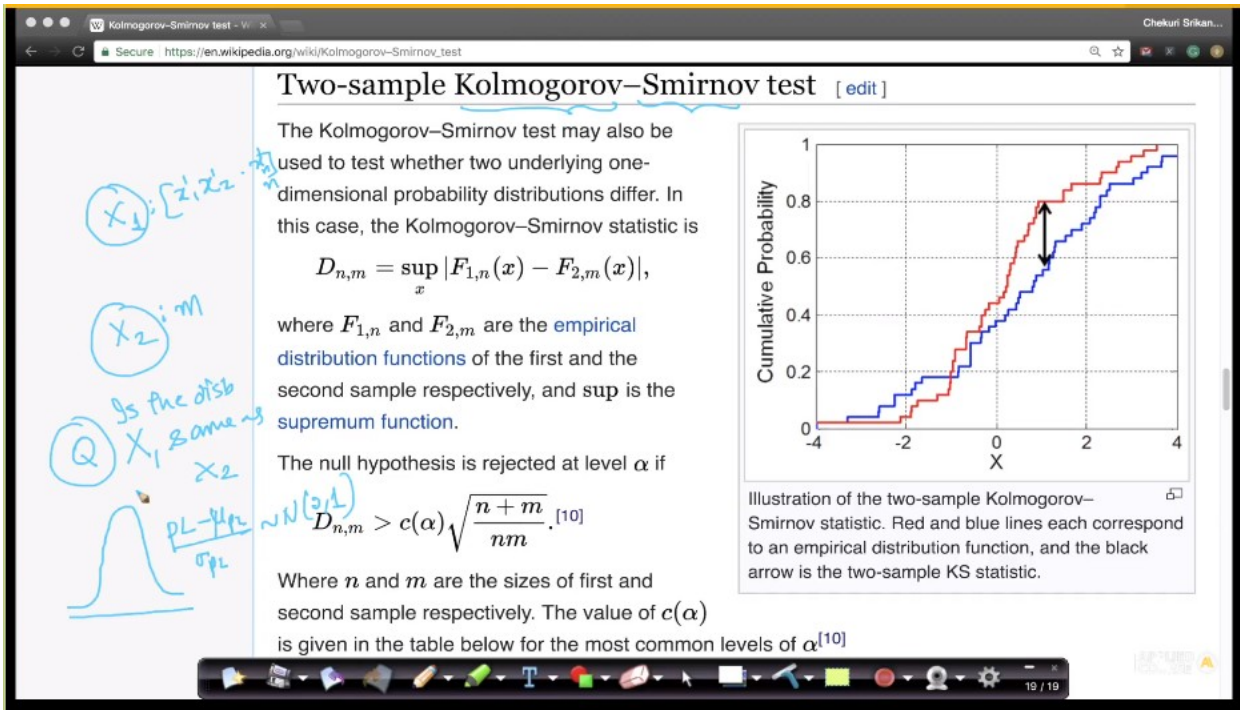
It gives **pvalue**, as the **pvalue is large** then the random variable is normal.

For uniform distribution:



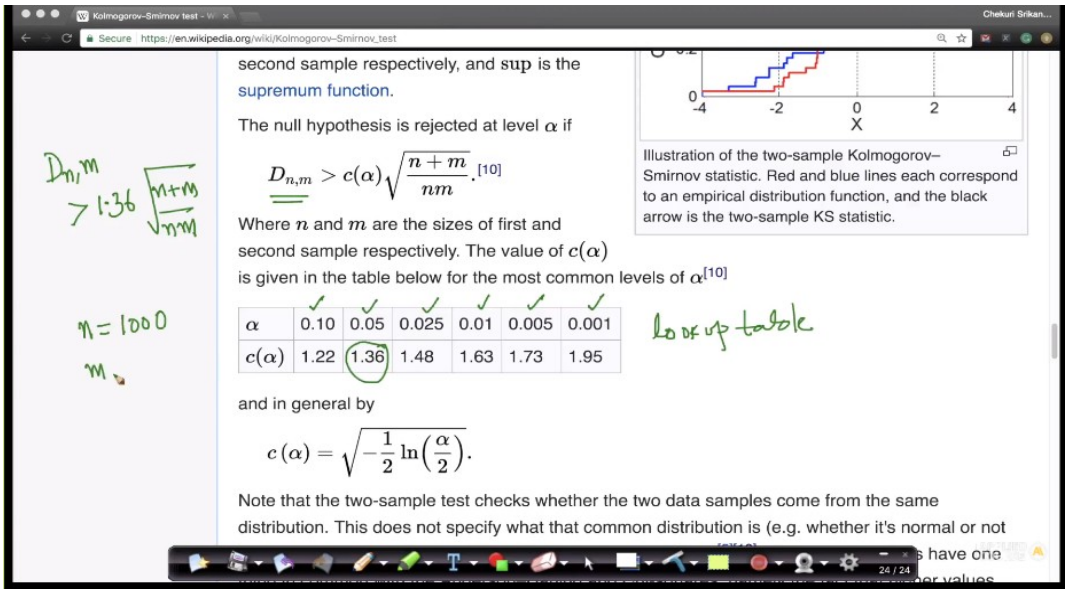
Steps for calculating K-S test similarity:

Standardize the data.



Here blue and red cdf's have m and n observations.

- Define null hypothesis, the two distributions are same.
- The sup is the maximum difference. Which is D(n,m).
- For rejecting null hypothesis we have a formula (2nd formula).



α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95