

## NOTES

Q-Q plot:

Step – 1

The  $x_1$  to  $x_{500}$  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.

Quantile-Quantile (Q-Q) plot :

$X: x_1, x_2, x_3, \dots, x_{500}$

(Q) is  $X$  Gaussian disb?  $\rightarrow$  Q-Q plot (Graphical)  
 $\rightarrow$  statistical testing (KS (A.D))

(How) ① Sort  $x_i$ 's & compute percentiles.

$x_1, x_2, \dots, x_{500}$   
 $\downarrow$  sort (asc.)  
 $x'_1, x'_2, \dots, x'_{500}$   $\xrightarrow{\text{percentiles}}$   $x^{(1)}_5, x^{(2)}_{10}, x^{(3)}_{15}, \dots, x^{(100)}_{500}$

$\rightarrow$  1st. percentile values of  $x_i$ 's  
 $x^{(1)}_5, x^{(2)}_{10}, x^{(3)}_{15}, \dots, x^{(100)}_{500}$

APPLIED COURSE

Step – 2

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

②  $Y \sim N(0, 1)$   $\rightarrow$  std. Normal disb

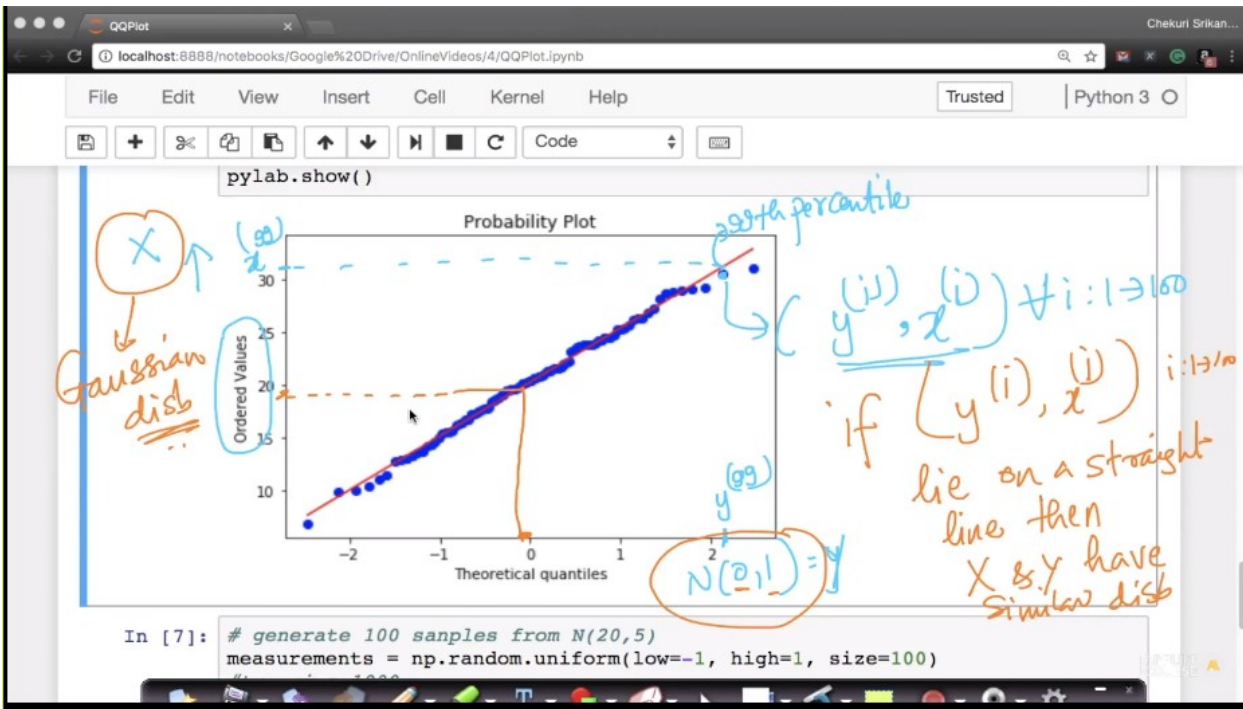
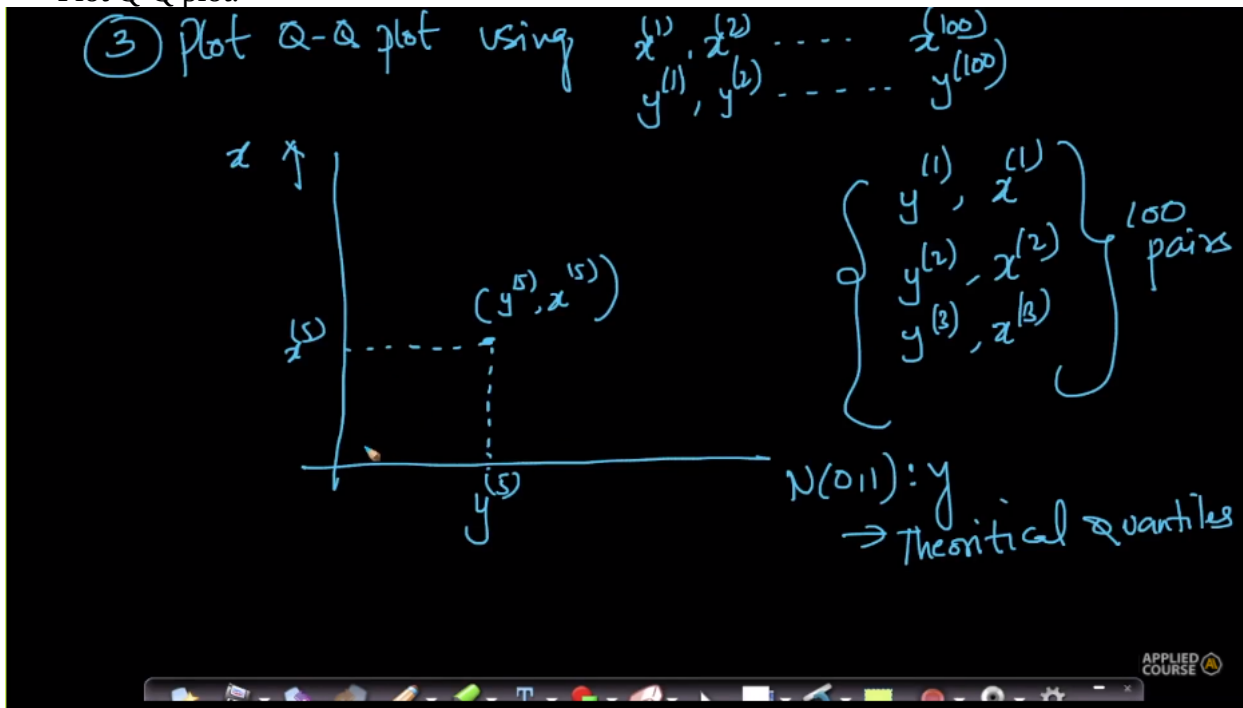
$y_1, y_2, y_3, \dots, y_{1000} \rightarrow 1000$  obs from  $N(0, 1)$

$\downarrow$  sort (asc.)  
 $y'_1, y'_2, y'_3, \dots, y'_{1000}$   
 $\downarrow$  percentiles  
 $y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}, \dots, y^{(100)}$

APPLIED COURSE

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-normal** distribution.

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

- Pareto to Gaussian.

Handwritten notes on a blackboard explaining the conversion of a Pareto distribution to a Gaussian distribution using the Box-Cox transformation.

Pareto  $\sim X: [x_1, x_2, \dots, x_n]$  → Conversion

Gaussian  $\sim Y: [y_1, y_2, \dots, y_n]$

①  $\text{box-cox}(X) = \lambda$  (labeled  $\alpha$  and  $\lambda$ )

②  $y_i = \begin{cases} \frac{x_i - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x_i) & \text{if } \lambda = 0 \end{cases}$

Handwritten notes on the right: if  $\lambda = 0 \Rightarrow x_i \sim \log\text{-normal}$  else

Handwritten notes on the left: Gaussian dist

Handwritten notes on the right: beyond the scope of this course

Screenshot of the SciPy documentation for `scipy.stats.boxcox`. The page shows the function signature, parameters, and returns. Handwritten notes in orange ink are present:

- $y = \text{boxcox}(x)$  with an arrow pointing to the `boxcox` function.
- A diagram showing a triangle with  $\lambda$  at the bottom and  $\Delta$  at the top, with an arrow pointing from the function to the triangle.

Documentation details:

- Function: `scipy.stats.boxcox(x, lmbda=None, alpha=None)`
- Return: a positive dataset transformed by a Box-Cox power transformation.
- Parameters:
  - `x`: ndarray. Input array. Should be 1-dimensional.
  - `lmbda`: {None, scalar}, optional. If `lmbda` is not None, do the transformation for that value. If `lmbda` is None, find the lambda that maximizes the log-likelihood function and return it as the second output argument.
  - `alpha`: {None, float}, optional. If `alpha` is not None, return the `100 * (1-alpha)%` confidence interval for `lmbda` as the third output argument. Must be between 0.0 and 1.0.
- Returns: `boxcox`: ndarray. Box-Cox power transformed array.

Pearsons correlation coefficient:

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

pearson correlation coefficient (pcc)

$$\rho_{x,y} = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$$

$$\sigma_x = \sqrt{\text{Var}(x)}$$

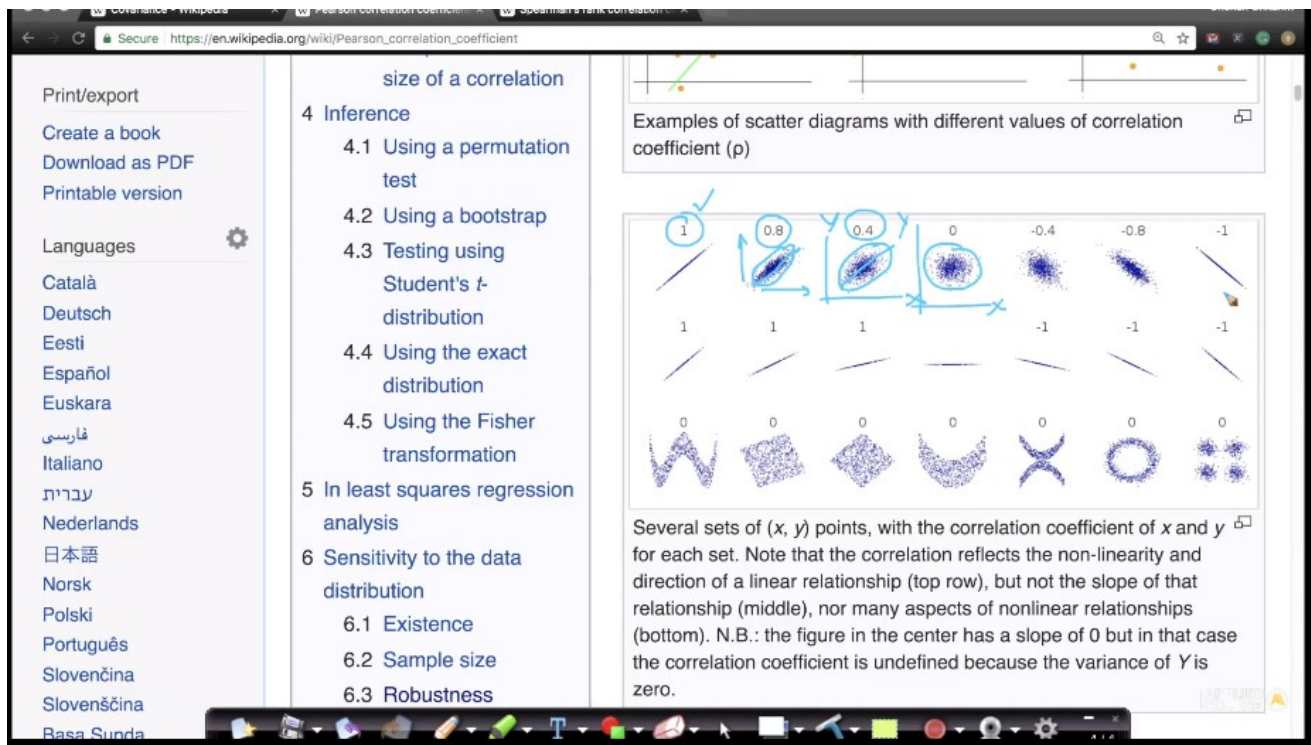
$x \uparrow, y \uparrow$  ,  $\text{Cov}(X,Y)$  (+ve)  
 $x \uparrow, y \downarrow$  ,  $\text{Cov}(X,Y)$  (-ve)

MORE VIDEOS

APPLIED COURSE

## Limitations:

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





## Spearman's rank – correlation coefficient:

Spearman rank – corr. coeff. ( $\gamma$ )

$\rho_{xy} \rightarrow$  linear relationship

$\gamma = \rho_{\underline{x}, \underline{y}}$

$\rho = 1 \leftarrow$  linear  $x \uparrow y \uparrow$

$\rho = -1 \leftarrow$  linear  $x \uparrow y \downarrow$

linear or not

linear or not

$\gamma = 1$

$\gamma = -1$

	$\checkmark x$	$\checkmark y$	$\textcircled{\gamma x}$	$\gamma y$
$s_1$	$\textcircled{160}$	52	4	3
$s_2$	150	66	2	4
$s_3$	170	68	5	5
$s_4$	140	46	1	1
$s_5$	158	51	3	2

APPLIED COURSE

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

Covariance - Wikipedia

Pearson correlation coefficient

Spearman's rank correlation coefficient

Contents [hide]

- 1 Definition and calculation
- 2 Related quantities
- 3 Interpretation
- 4 Example
- 5 Determining significance
- 6 Correspondence analysis based on Spearman's rho
- 7 See also
- 8 References
- 9 Further reading
- 10 External links

Definition and calculation [edit]

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables [3]

when the data are roughly elliptically distributed and there are no prominent outliers, the Spearman correlation and Pearson correlation give similar values.

Spearman correlation = 0.84

Pearson correlation = 0.67

The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in

Confidence intervals:

## point estimate:

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

Confidence Interval:

disb  $\leftarrow X$ : heights  
 $\{x_1, x_2, x_3, \dots, x_{10}\}$  - random sample from  $X$  of size 10

estimate the population mean of  $X = \mu$

$\mu \approx \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  - simple avg  
pop-mean  $\rightarrow$  sample-mean  
 $\rightarrow$  as  $n \uparrow$ ,  $(\bar{x} \rightarrow \mu)$

$\mu \approx \bar{x}$   
point estimate

APPLIED COURSE

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:

$\{x_1, x_2, \dots, x_{10}\}$   
 $\{180, 162, 158, 172, 168, 150, 171, 183, 165, 176\}$   $\rightarrow$  heights of people in cm

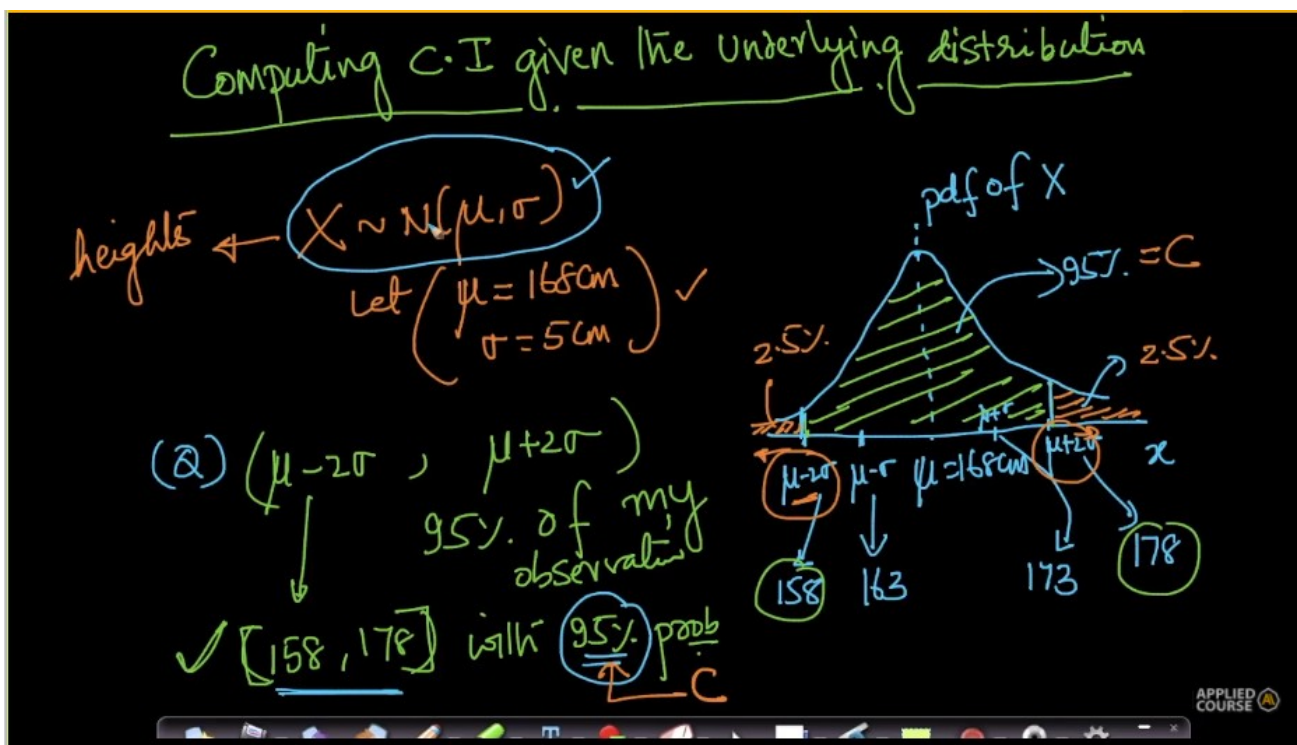
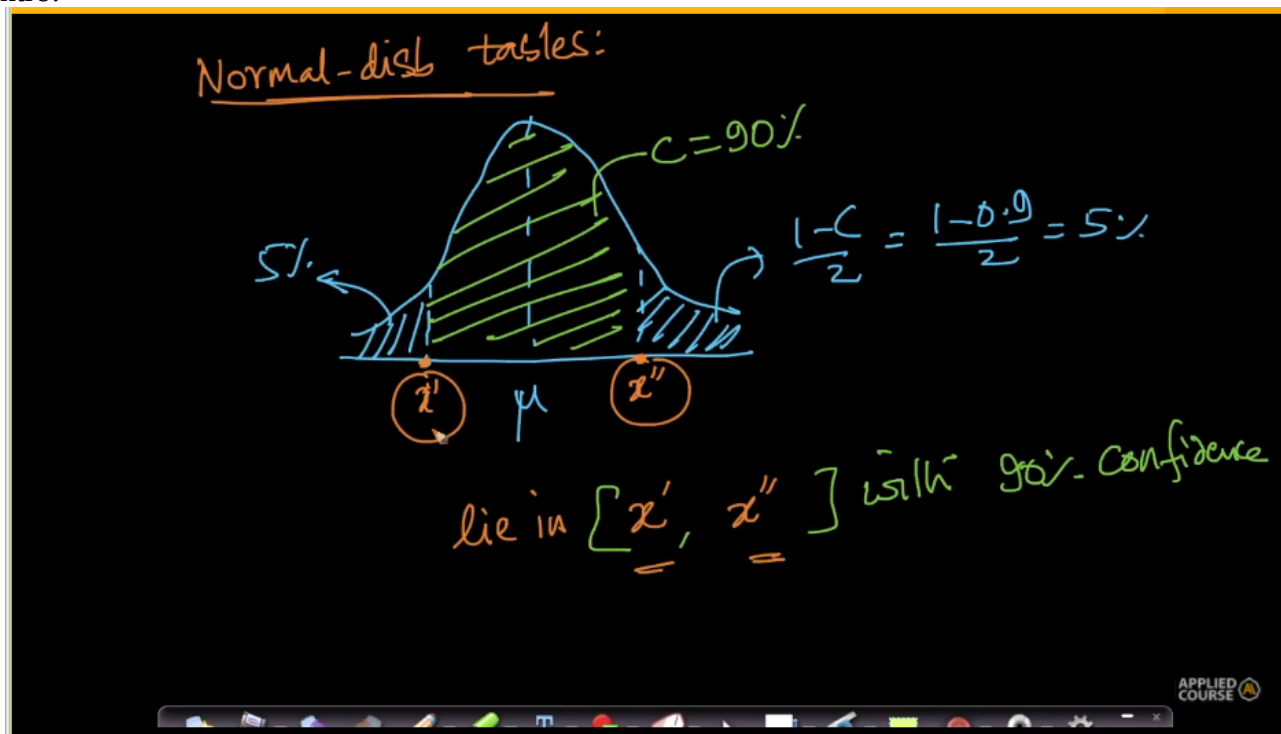
POINT ESTIMATE of  $\mu = \frac{1}{10} \sum_{i=1}^{10} x_i = 168.5 \text{ cm}$  ✓

$\mu \in [162.1, 174.9]$  with 95% probability  
pop-mean  $\rightarrow$  interval  $\rightarrow$  Confidence

C.I. ✓

APPLIED COURSE

Calculation of confidence intervals:  
Intro:



calculation:

C.I for mean( $\mu$ ) of a r.v

$X \sim \underline{F}$  with pop-mean of  $\mu$  & std-dev of  $\sigma$

$\downarrow \{x_1, x_2, \dots, x_{10}\} \rightarrow$  Sample of size  $n=10$

$\checkmark \{180, 162, 158, 172, 168, 150, 171, 183, 165, 176\}$  - given this sample

(Q) What is the 95% C.I of  $\mu$

APPLIED COURSE

Central limit theorem:

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

Case – 1(you have  $\sigma$  and  $\mu$  of population):

Case 1:  $\sigma = 5\text{cm}$  {we know pop-std-dev}

CLT:

$\bar{x} = \text{Sample mean} = \frac{1}{10} \sum_{i=1}^{10} x_i$  ( $n=10$ )

$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$  (CLT)

Sample-mean  $\rightarrow \bar{x}$

pop-mean  $\rightarrow \mu$

pop-std-dev  $\rightarrow \frac{\sigma}{\sqrt{n}}$

$\left\{ \mu \in \left[ \bar{x} - \frac{2\sigma}{\sqrt{n}}, \bar{x} + \frac{2\sigma}{\sqrt{n}} \right] \right\}$  with 95% Confidence

$\mu - 2\sigma$   $\mu + 2\sigma$

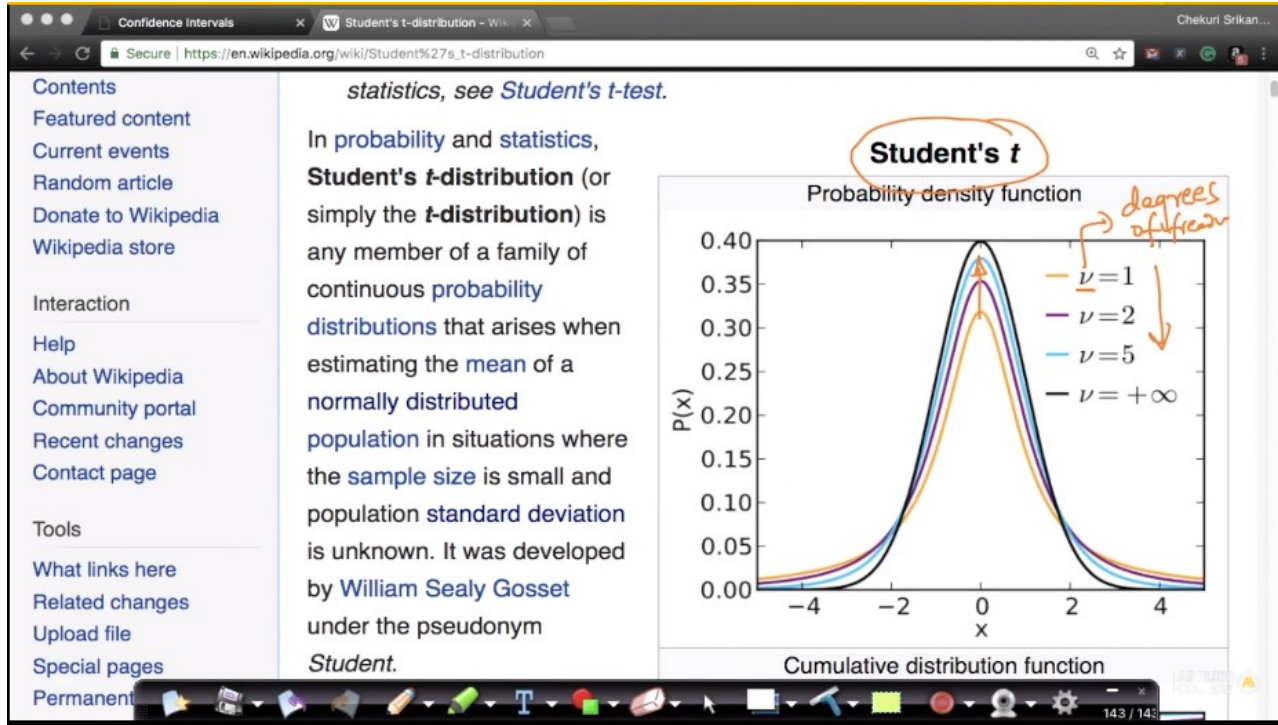
APPLIED COURSE



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

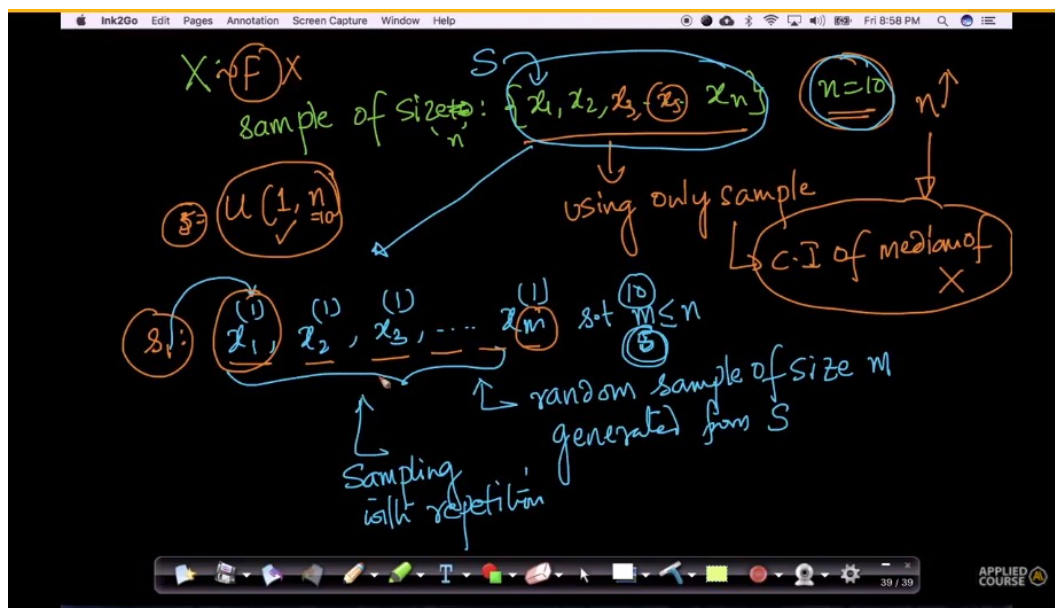
### Case – 2:(If we do not know population standard deviation)

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

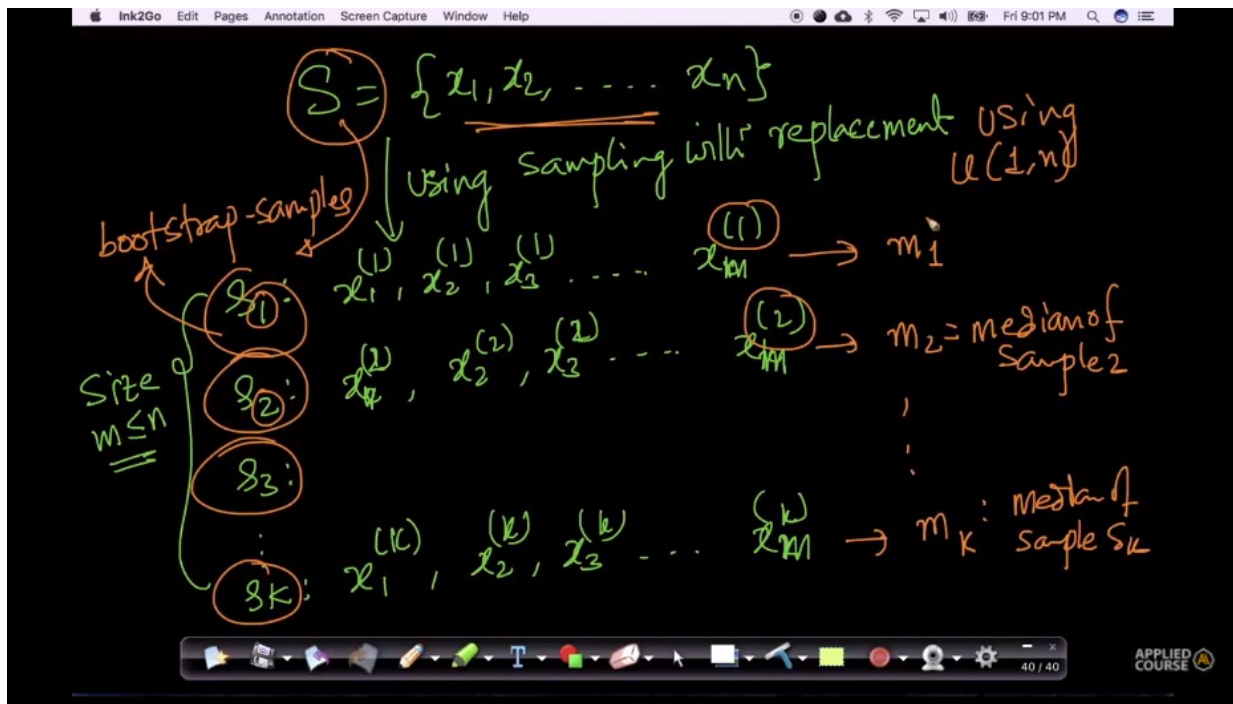


### Confidence interval using bootstrapping:

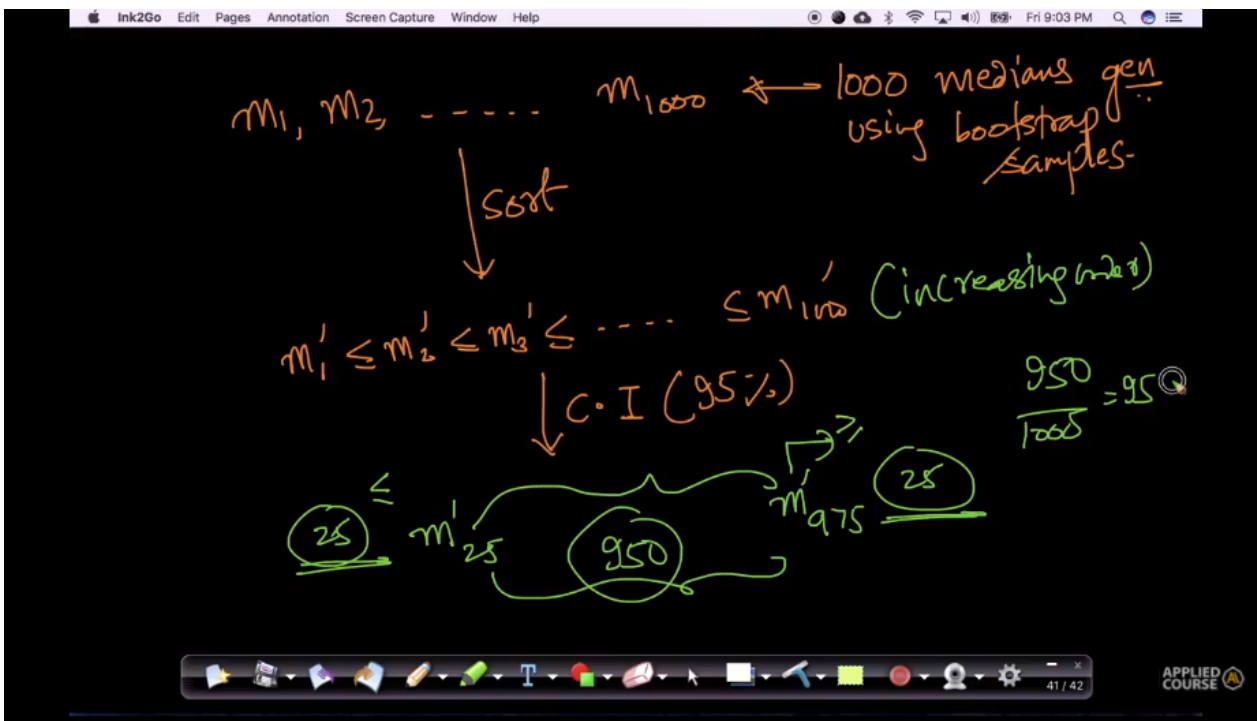
- The sampling is done by replacement, for choosing the values we can use uniform distribution.



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



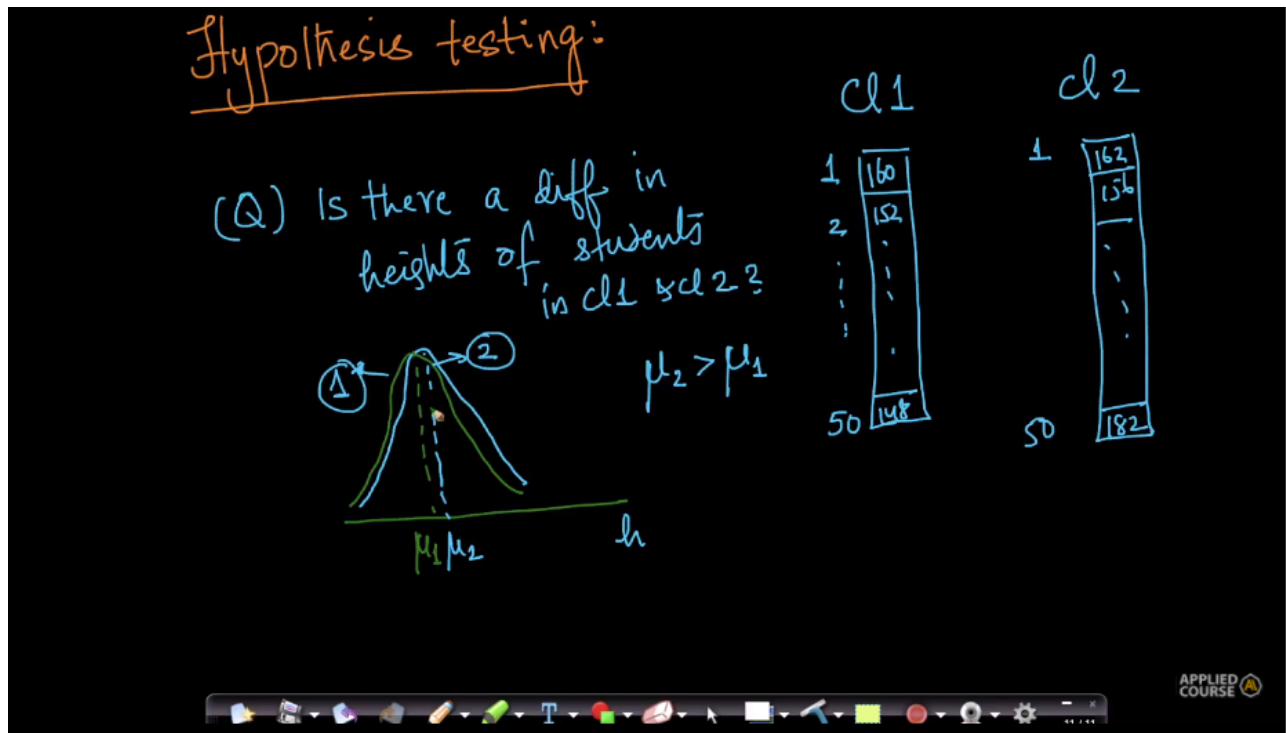
- For calculation of median of the sample, first the medians of the bootstrap samples are made.
  - Then the medians are sorted in increasing order.
- Then the confidence interval is calculated, here m25 and m975 are the 95 percent CI for median of the sample.



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:

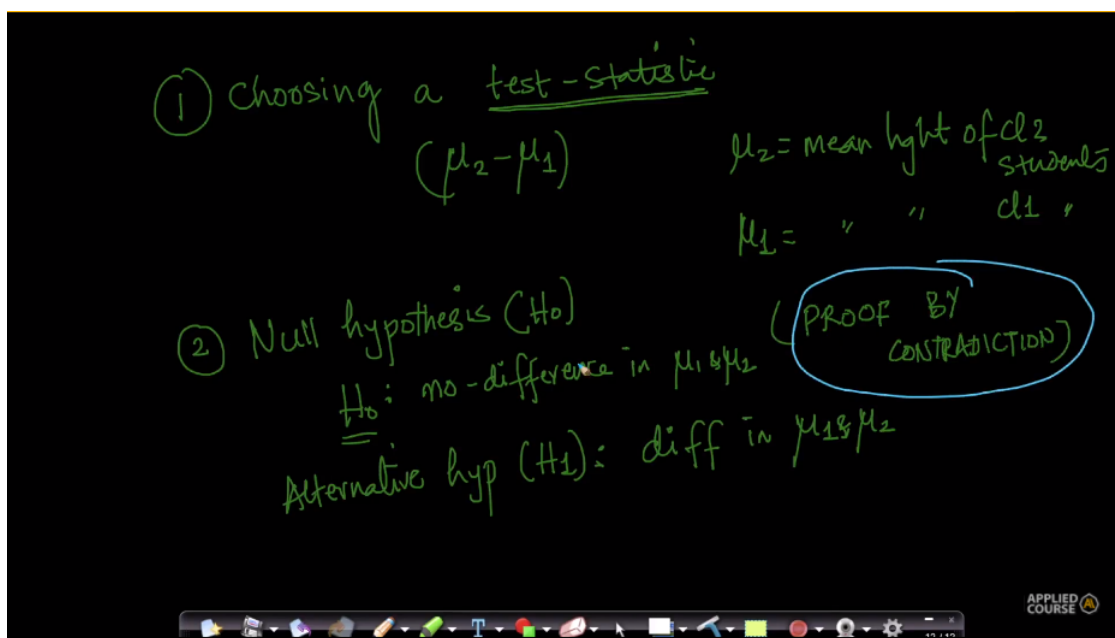


Step – 1

Choosing the test – statistic.

Step – 2

Choosing null – hypothesis.



Step – 3

Computing the p – value.

③ p-value: prob. of obs  $(\mu_2 - \mu_1)$  if null hyp is true.

assume  $H_0$  is true.

accept  $H_0$  ← if p-value = 0.9 ✓

⇒ prob of 10CM is 0.9 if  $H_0$  is true

reject  $H_0$  ← if p-value = 0.05 → 5% chance that 10CM if  $H_0$  is true

cl1 cl2  
✓50 ✓50

APPLIED COURSE

How to compute p – value:

- Step – 1

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

p-value: (permutations testing)

cl1 cl2

1 1  
50 50

$\mu_1$   $\mu_2$  →  $\Delta$

Jumble

randomly sample

10k

X Y

1 1  
50 50

$\mu_1 - \mu_2 \rightarrow S_1$

2 1 1  
50 50

$\mu_1 - \mu_2 \rightarrow S_2$

...

10k  $\rightarrow S_{10k}$

each time I sample 50 points I took the difference in

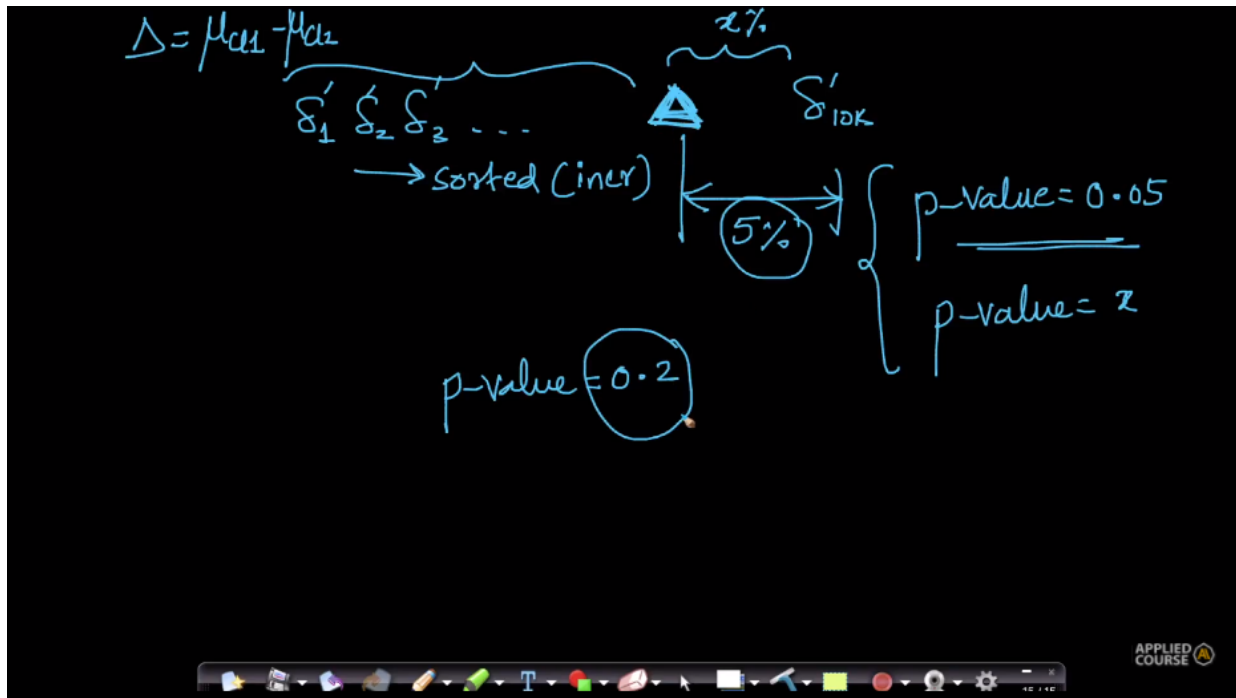
APPLIED COURSE



- Step – 2

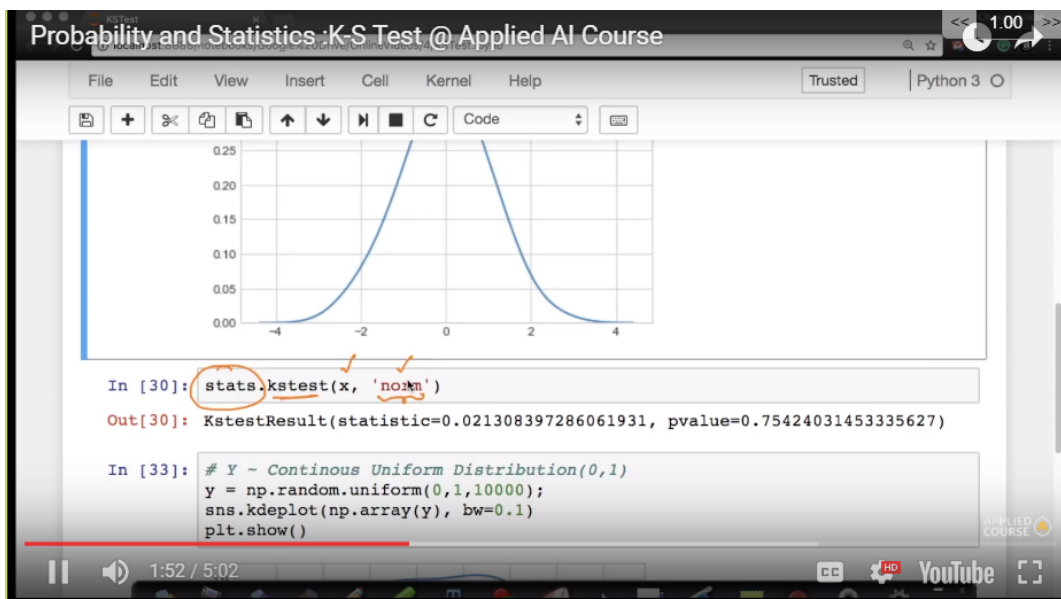
Then sort the difference of the **mean statistic** that is chosen from the sampled classes. That is all **delta's**.

- Step – 3



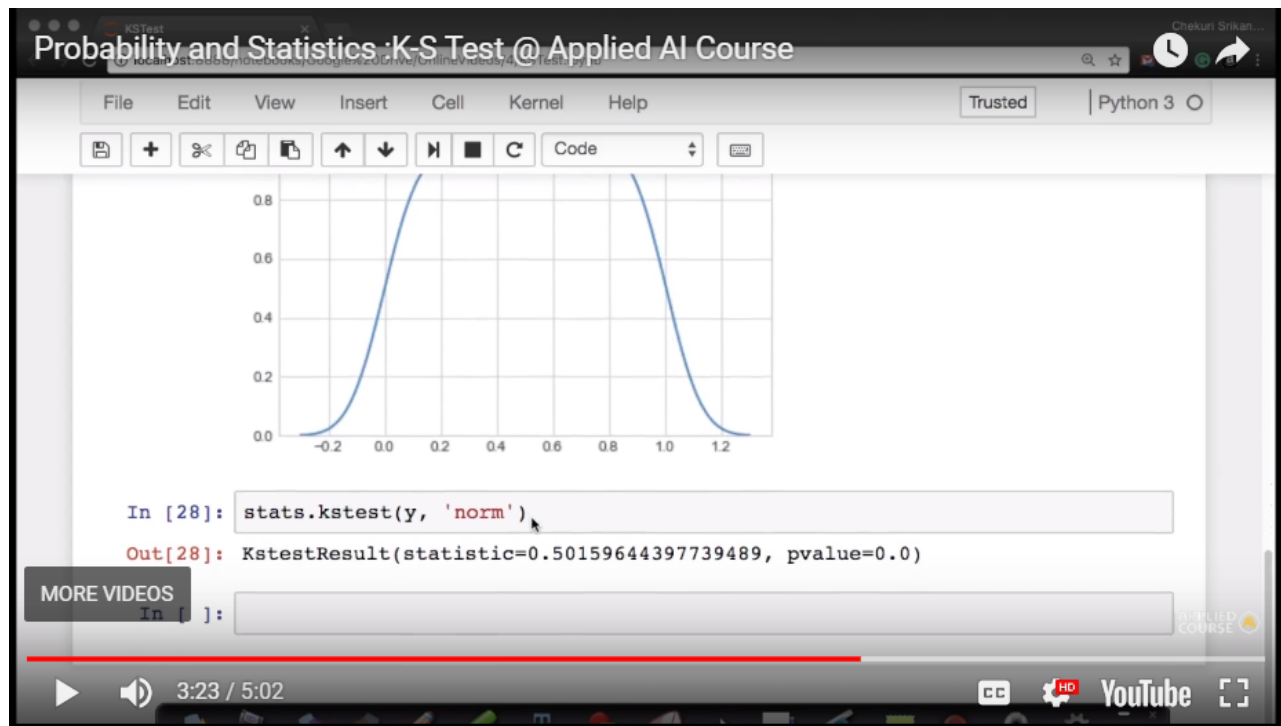
## K-S test:

This test is for similarity 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.

**Two-sample Kolmogorov–Smirnov test** [edit]

The Kolmogorov–Smirnov test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

where  $F_{1,n}$  and  $F_{2,m}$  are the empirical distribution functions of the first and the second sample respectively, and  $\sup$  is the supremum function.

The null hypothesis is rejected at level  $\alpha$  if

$$D_{n,m} > c(\alpha) \sqrt{\frac{n+m}{nm}}.^{[10]}$$

Where  $n$  and  $m$  are the sizes of first and second sample respectively. The value of  $c(\alpha)$  is given in the table below for the most common levels of  $\alpha$ ^{[10]}

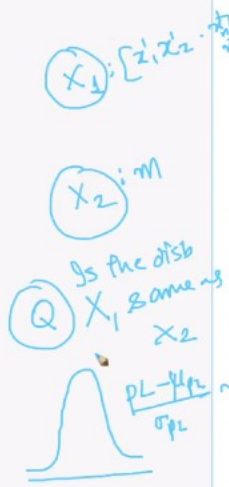
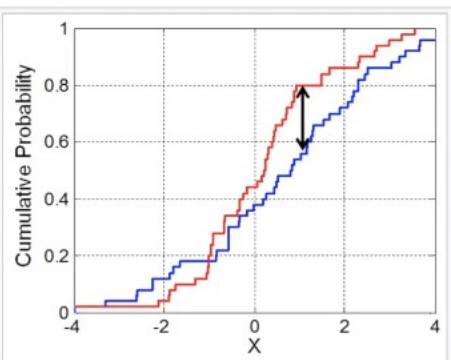



Illustration of the two-sample Kolmogorov–Smirnov statistic. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

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 (2<sup>nd</sup> formula).

second sample respectively, and  $\sup$  is the supremum function.

The null hypothesis is rejected at level  $\alpha$  if

$$D_{n,m} > c(\alpha) \sqrt{\frac{n+m}{nm}}.^{[10]}$$

Where  $n$  and  $m$  are the sizes of first and second sample respectively. The value of  $c(\alpha)$  is given in the table below for the most common levels of  $\alpha$ ^{[10]}

$\alpha$	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

and in general by

$$c(\alpha) = \sqrt{-\frac{1}{2} \ln\left(\frac{\alpha}{2}\right)}.$$

Note that the two-sample test checks whether the two data samples come from the same distribution. This does not specify what that common distribution is (e.g. whether it's normal or not)

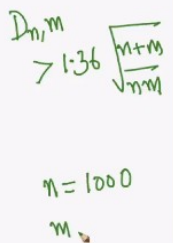
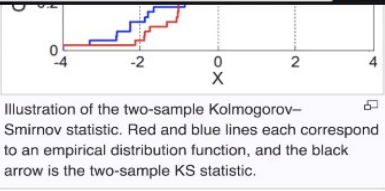



Illustration of the two-sample Kolmogorov–Smirnov statistic. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

look up table