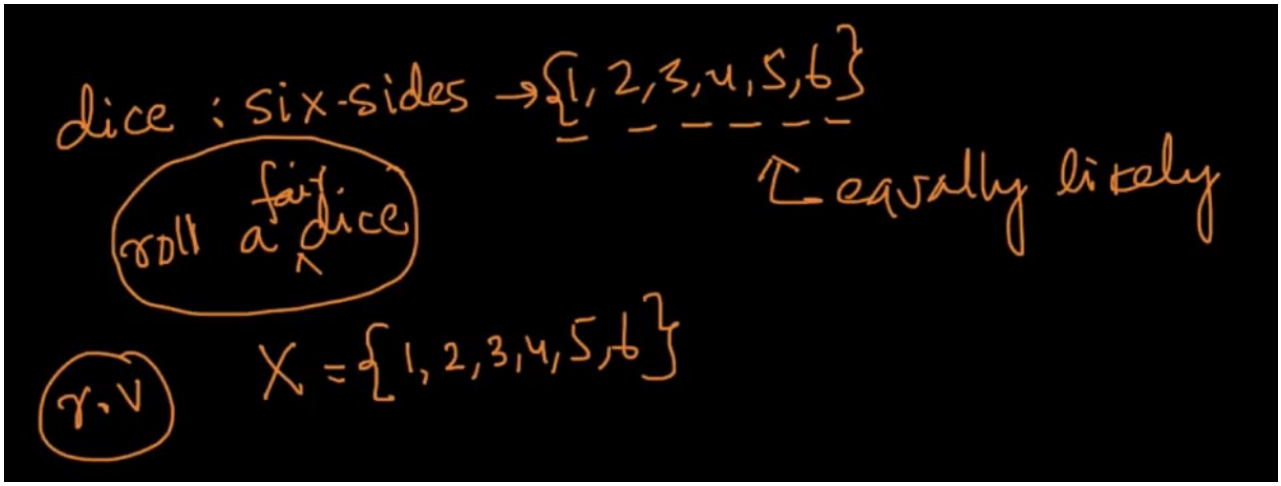
PROBABILITY AND STATISTICS

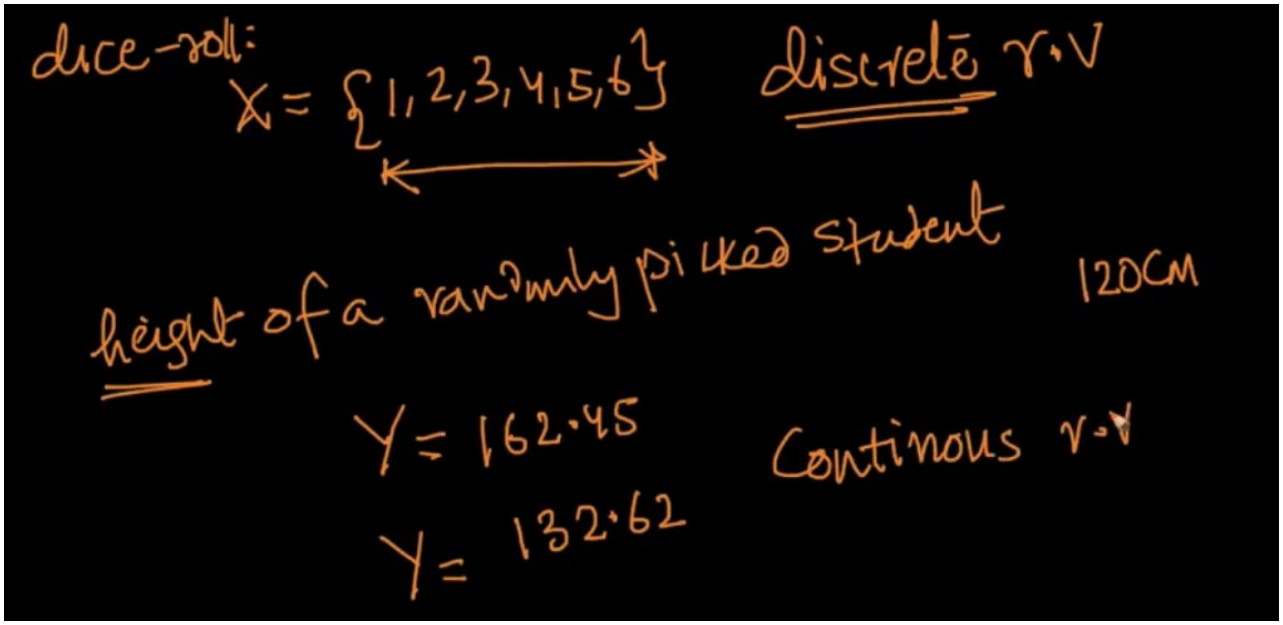
Basic Idea

Suppose, we've a new flower x_q in the dataset and we want to identify it's type.We can see that.It's lying in the intersection so we will give a percentage mentioned above i.e 80% chance it's versicolor.

Random Variable : X, is a variable whose possible values are numerical outcomes of a random phenomenon

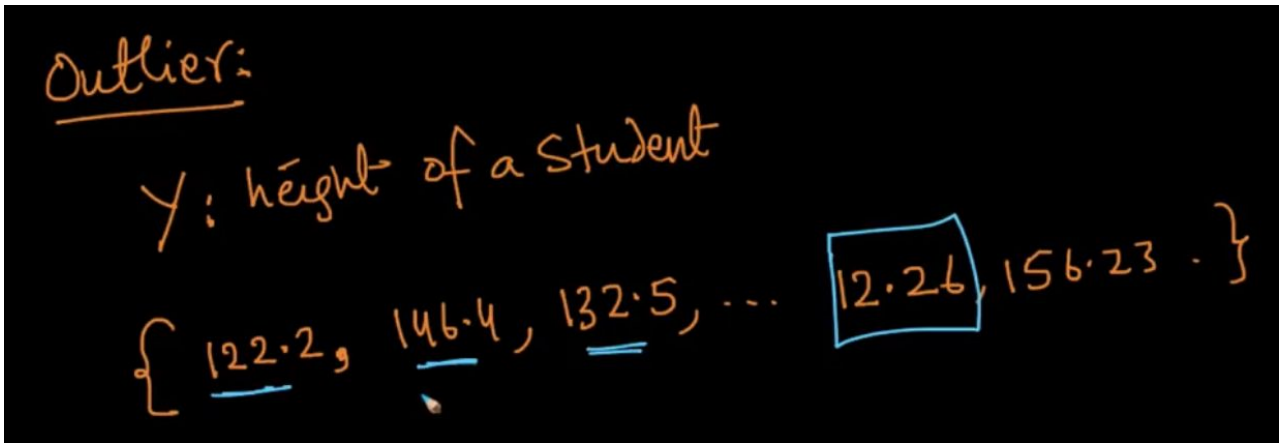


Types of Random variables

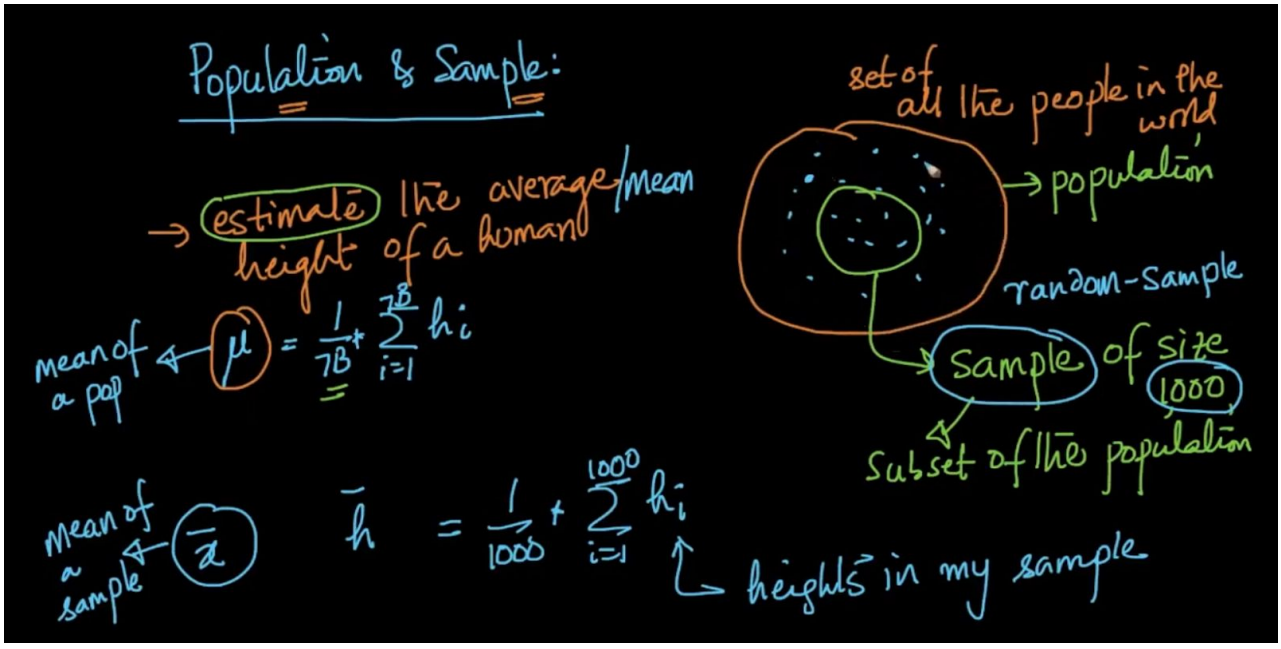


X is discrete r.v which can take value only from from a distinct set of points(ex: 4 not 4.5) while Y is continuous r.v which can take any real value (ex : any value)

OUTLIER

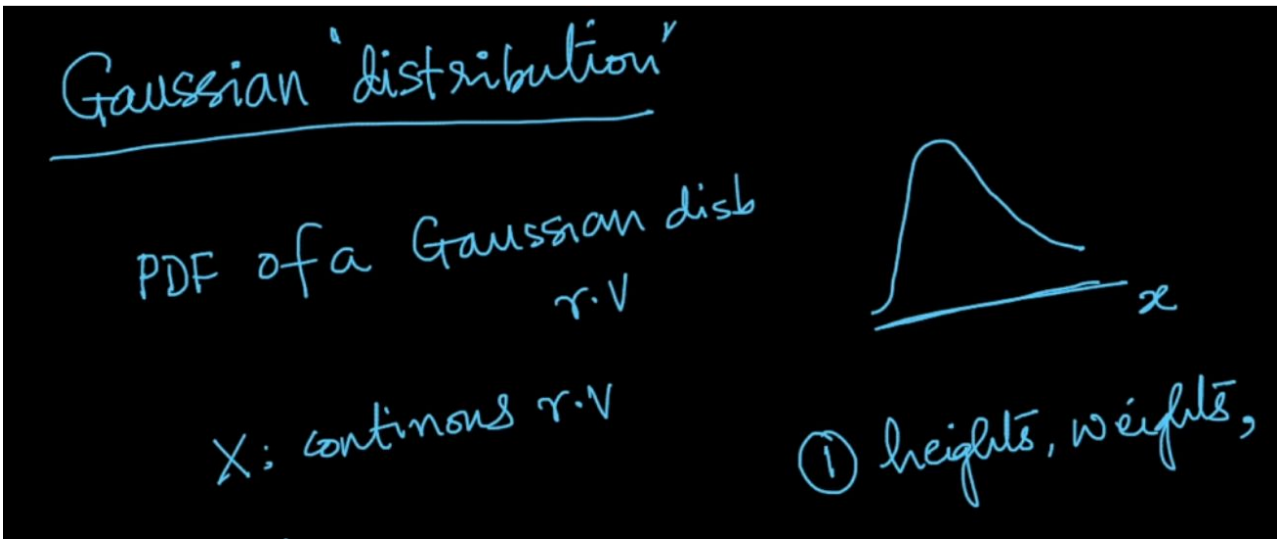


POPULATION AND SAMPLE



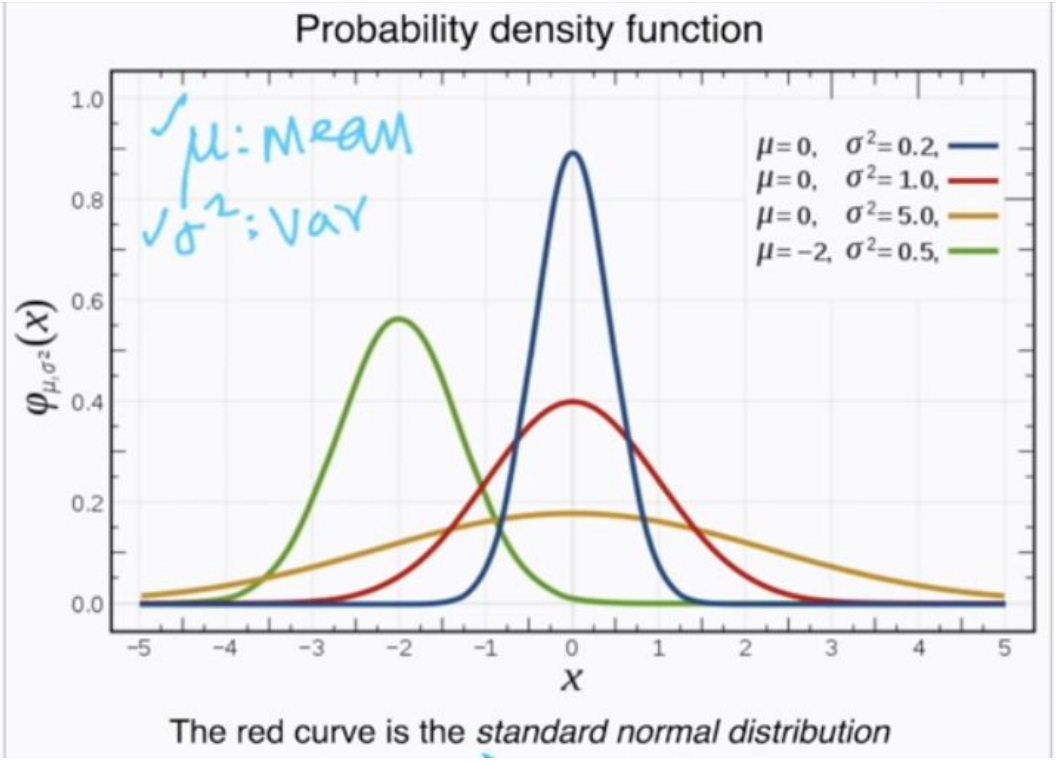
Suppose you want to take the average height of all the people in the world of 7B people. Theoretically you cant do that so you take a random sample of 1000 people based on the proportion they are distributed in 7B otherwise you might take just 1000 Indians or Americans which can do loss. As the sample (for eg: 1M from 1000) increases we get closer to the accurate answer

GAUSSIAN DISTRIBUTION

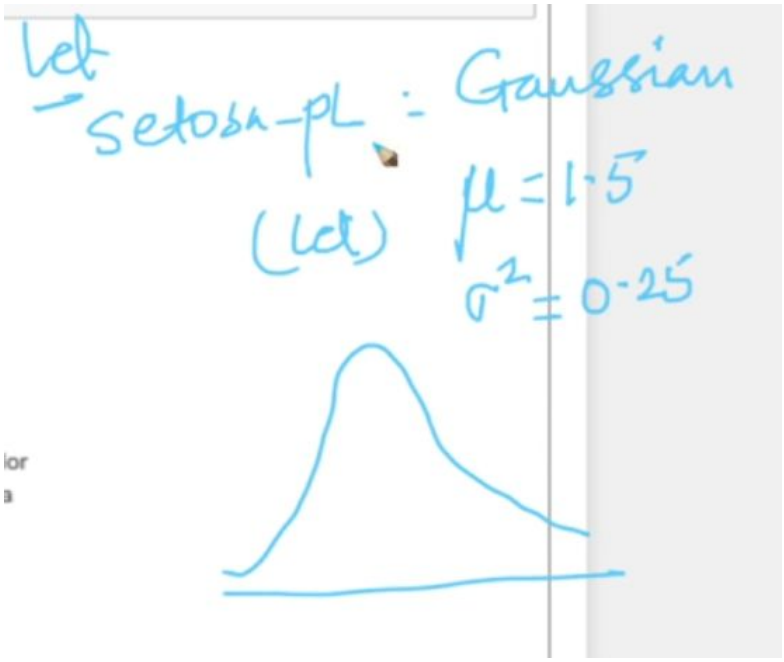


The bell-shaped curve is the PDF (Probability Distribution function) of Gaussian distributed random variable X

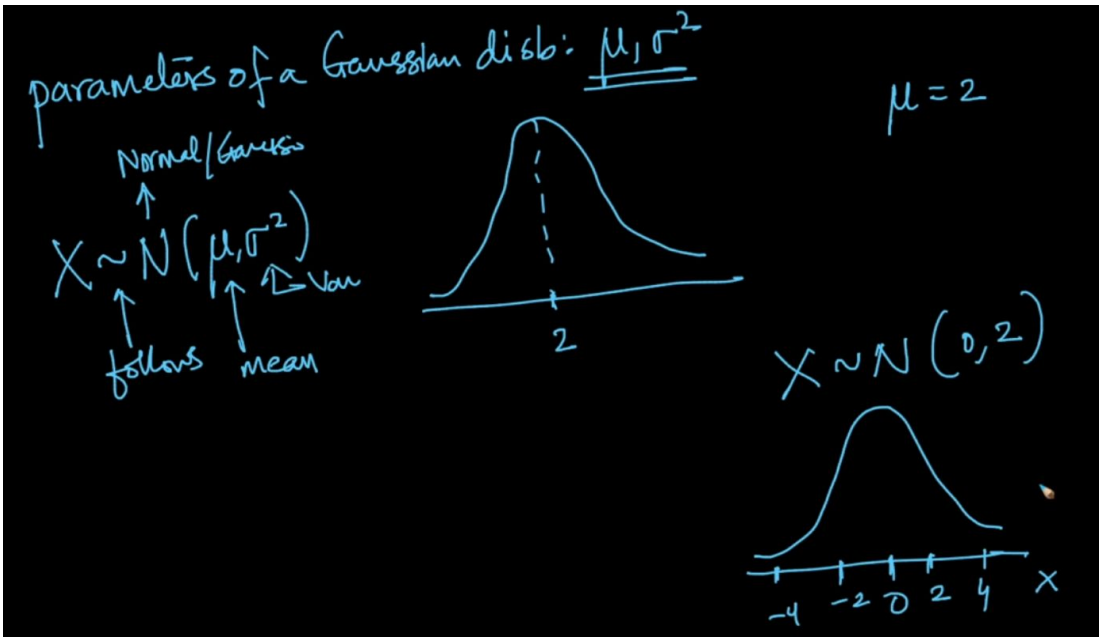
It can be used to find the distribution of weights ,heights etc.



If we've the mean (μ) and variance (σ^2) of the continuous random variable X. We can have the PDF of X. (μ, σ^2) are the parameters of the distribution function.

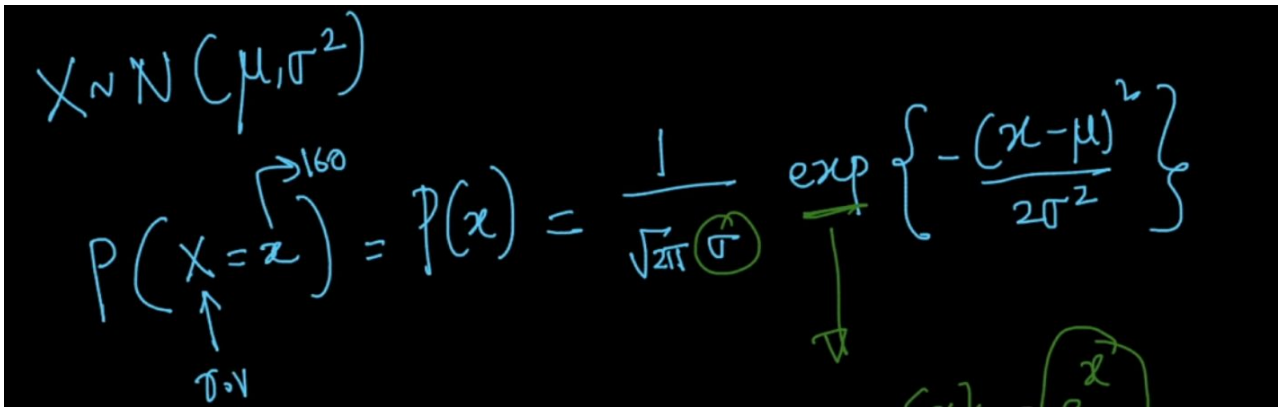


If the c.r.v[1] follows Gaussian then just by knowing its μ *and* σ^2 we can draw its distribution

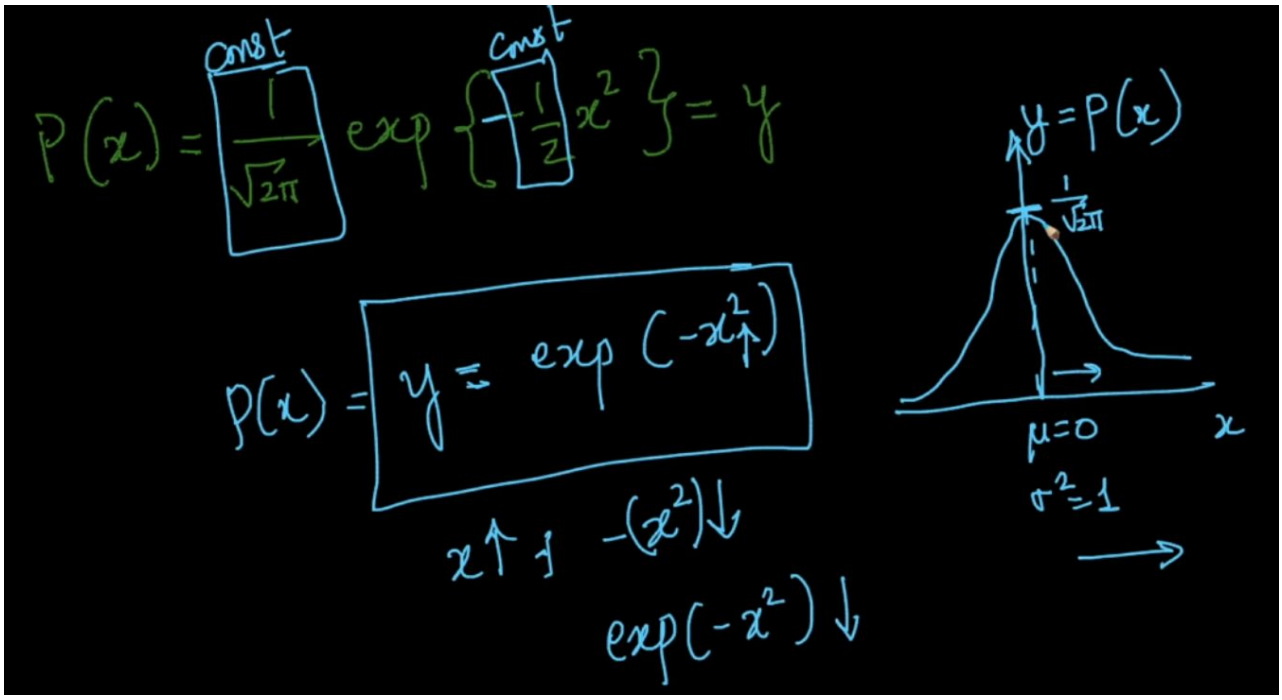


[1] C.r.v - Continuous random variable

Gaussian Mathematics

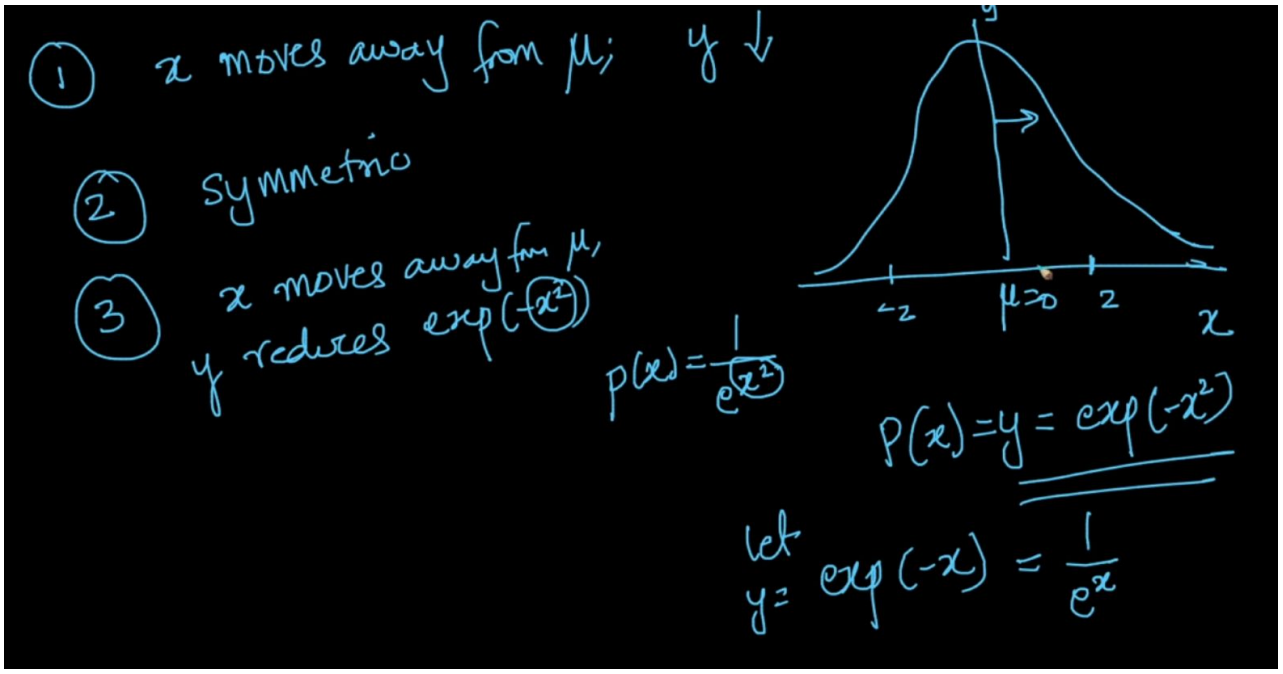


Breaking down of above formula. Let $\mu = 0$ *and* $\sigma = 1$

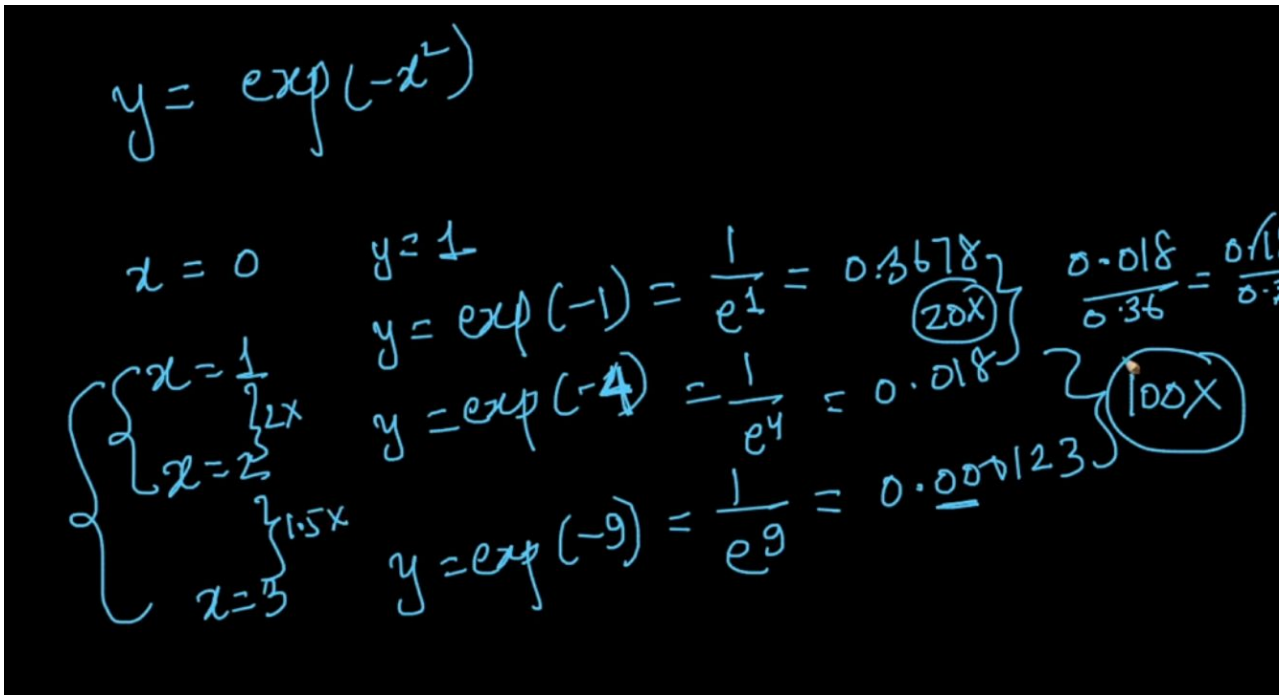


Here the above plot is at x = 0 so the max value ($y = \frac{1}{\sqrt{2\Pi}}$) and as the value of x increases the graph decreases exponentially squared ($y = \exp(-x^2) = \frac{1}{e^{x^2}}$ or $1 / e^{x^2}$)

Take - aways

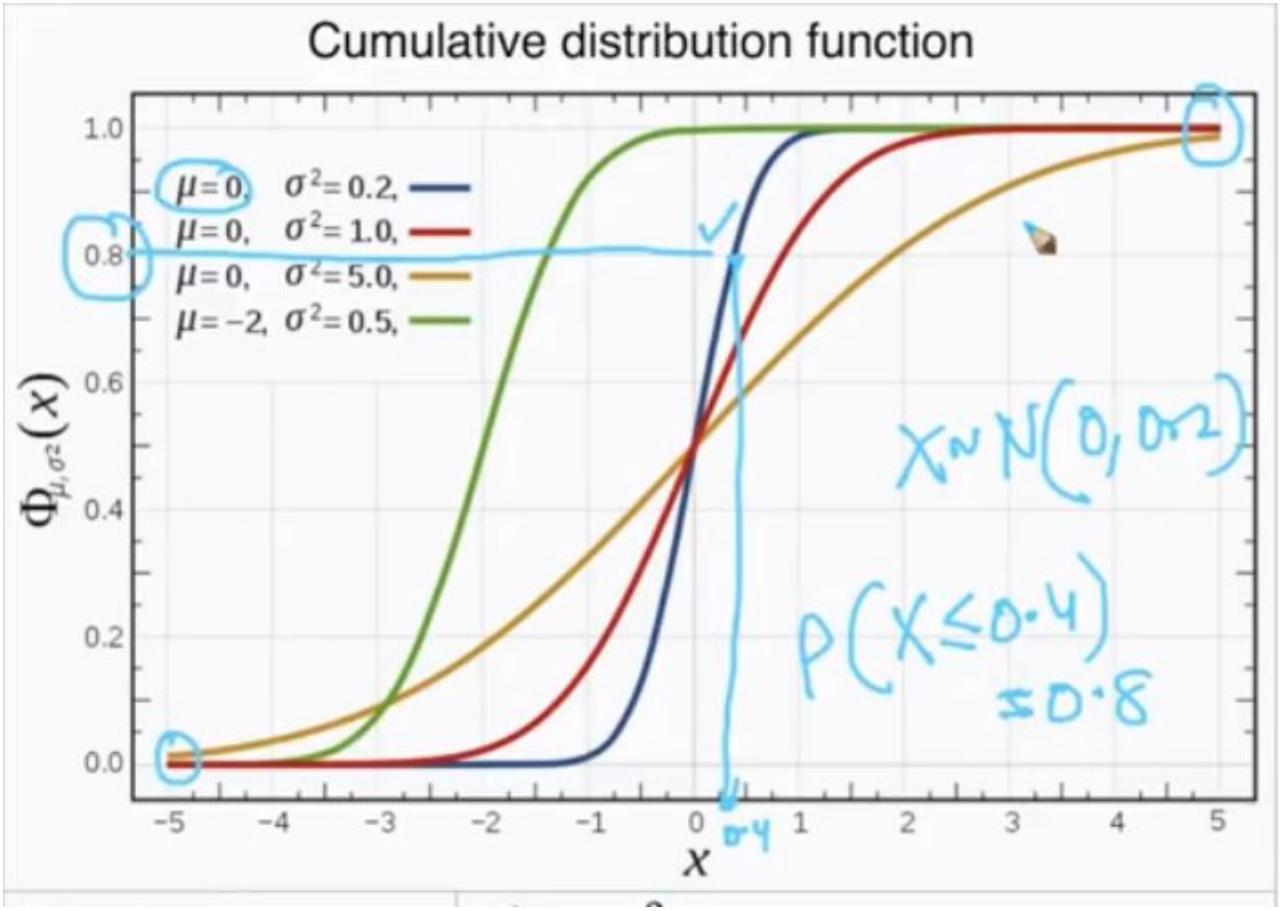


Let us discuss the 3rd point of exponential reduce.



As we can see the the x is increased just by 2x or 1.5x but the y is decreasing at 20x and 100x respectively. Hence , the bell curve

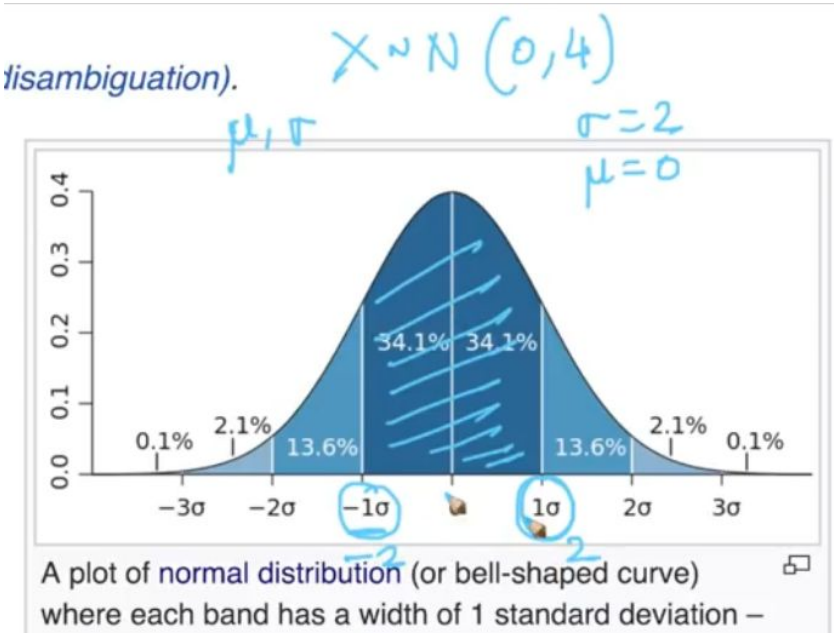
CUMULATIVE DISTRIBUTION FUNCTION



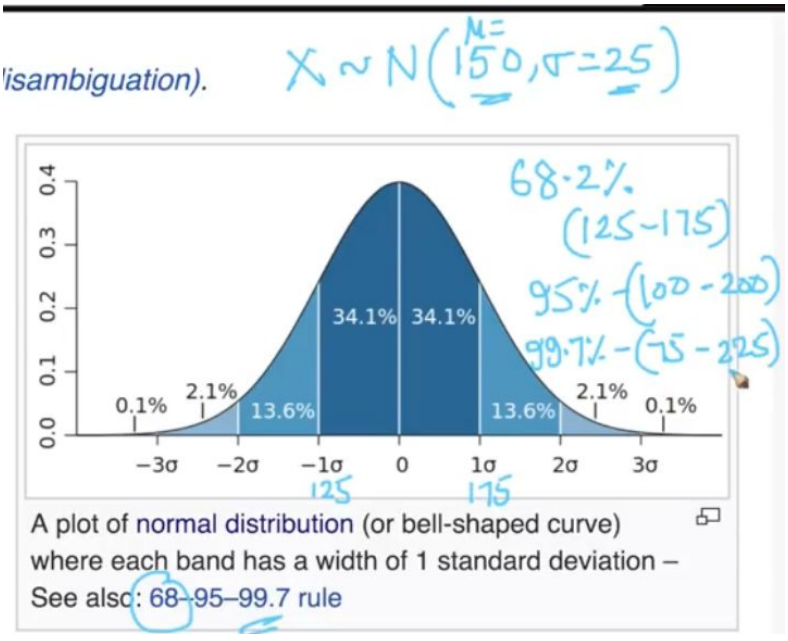
The height can tell us that P(X <= 0.4) = 0.8 (i.e it tells us about the probability that our r.v X takes the value <= 0.4 which is 0.8)

Takeaways from CDF :

1. Half of the values lie in X < 0 and y < 0.5 and half on X > 0 and y > 0.5
2. If variance is small then it is near X = 0 (blue) if it is large then it is farther(yellow)



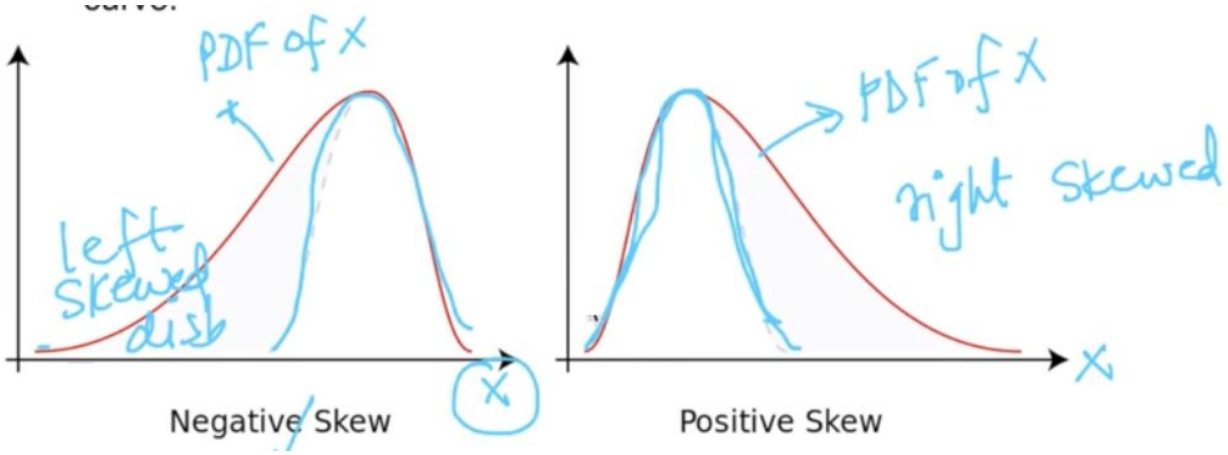
If $\mu = 0$ *and* $\sigma = 2$ then 68% are between -2 and 2 std.dev . We can get the % even without knowing about all the points. Let's take an example of heights of people below.



If $\mu = 150$ *and* $\sigma = 25$ of heights of peeps we can infer that 68% lie between (125 -175) and 95% between (100 - 200).

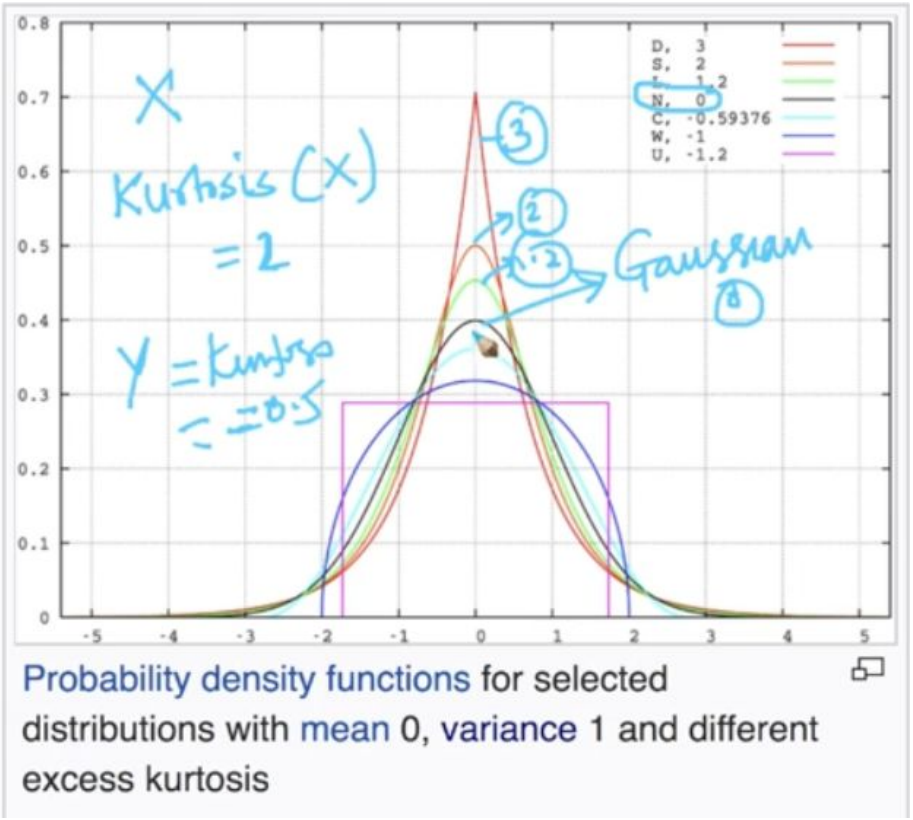
Standard Deviation ,68-95-99.7

SKEWNESS



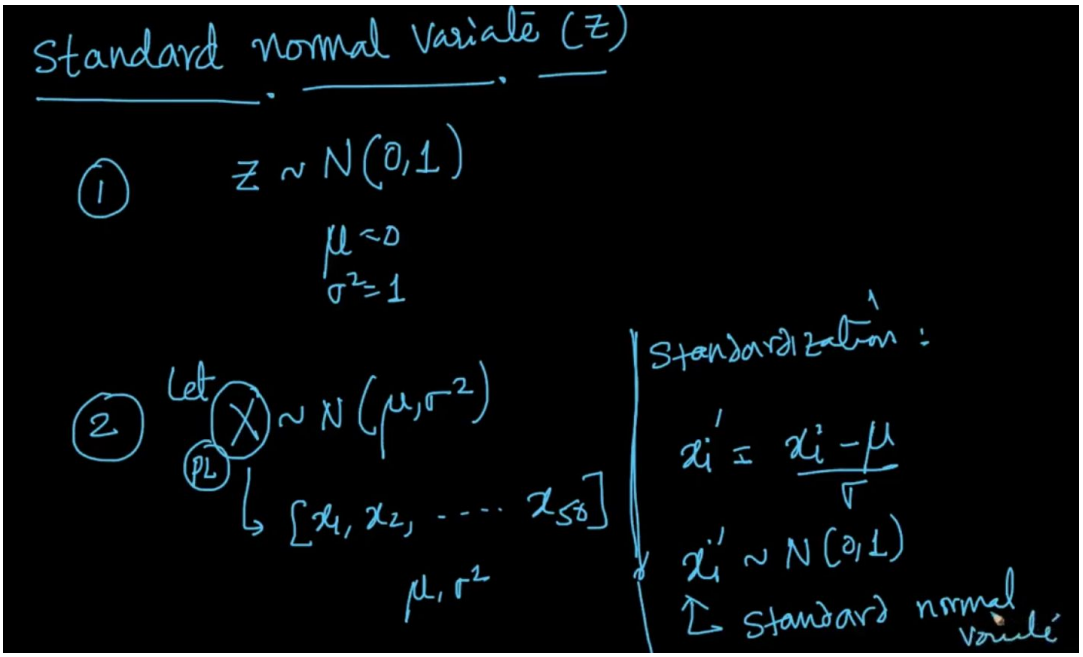
Skewness tells us about how far our distribution is from symmetric distribution. Skewness

KURTOSIS



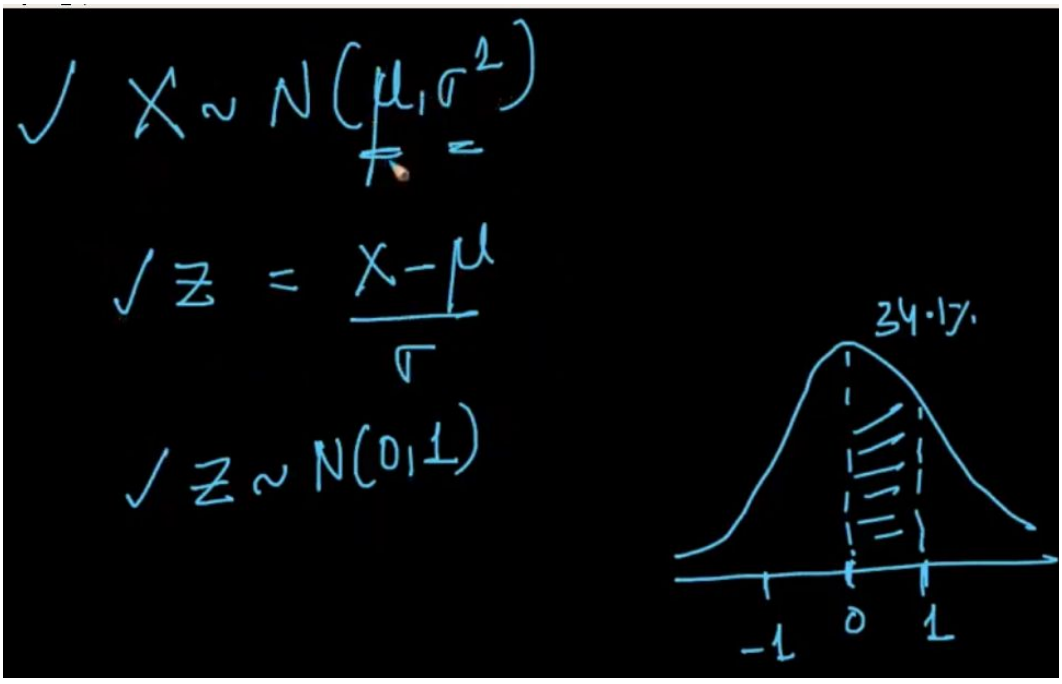
It measures how peaked your distribution is. So if Kurtosis (X) = 2 its peak will be greater than normal distribution and if negative it'll be lower. Kurtosis

STANDARD NORMAL VARIATE (*Z*)



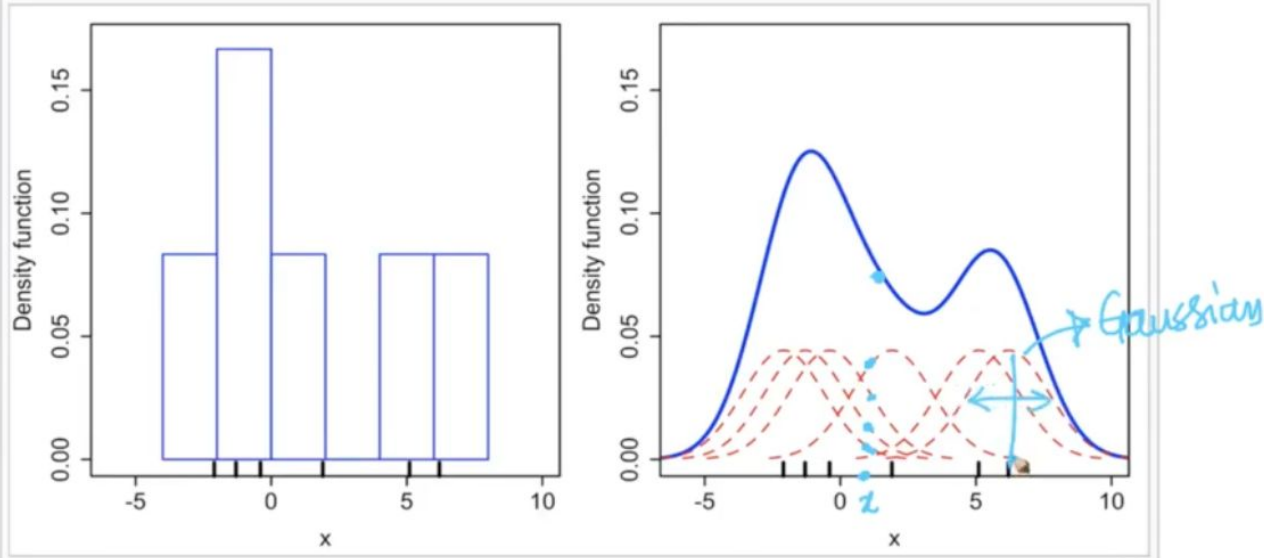
If you've a random variable **Z** such that its $\mu = 0$ *and* $\sigma^2 = 1$ then it is SNV

STANDARIZATION :



If we apply the above formula then we can get $Z \sim N(0,1)$. We can know that 68% of our data is between (-1,1) and 95% is between (-2,2) . We can transform any random variables to this.

KERNEL DENSITY ESTIMATION

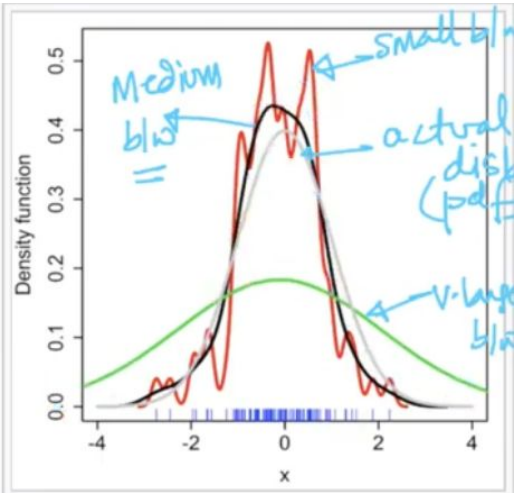


Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data. The 6 individual kernels are the red dashed curves, the kernel density estimate the blue curves. The data points are the rug plot on the horizontal axis.

Kernel density estimation we make graph in red(kernel) from the sample point and then we sum up all the heights as seen from x in the above diagram

BANDWIDTH : If bandwidth is low we get smoothed (red) if too large then too smoothed(green)

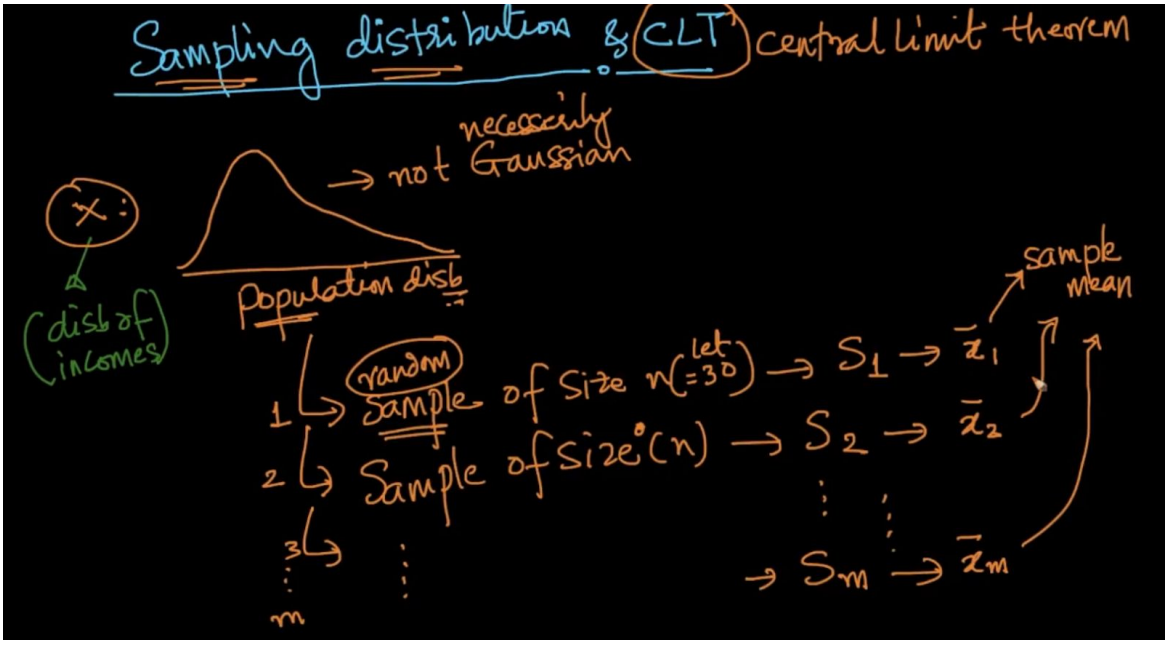
Kernel density estimation , Kernel Density Estimation intuit



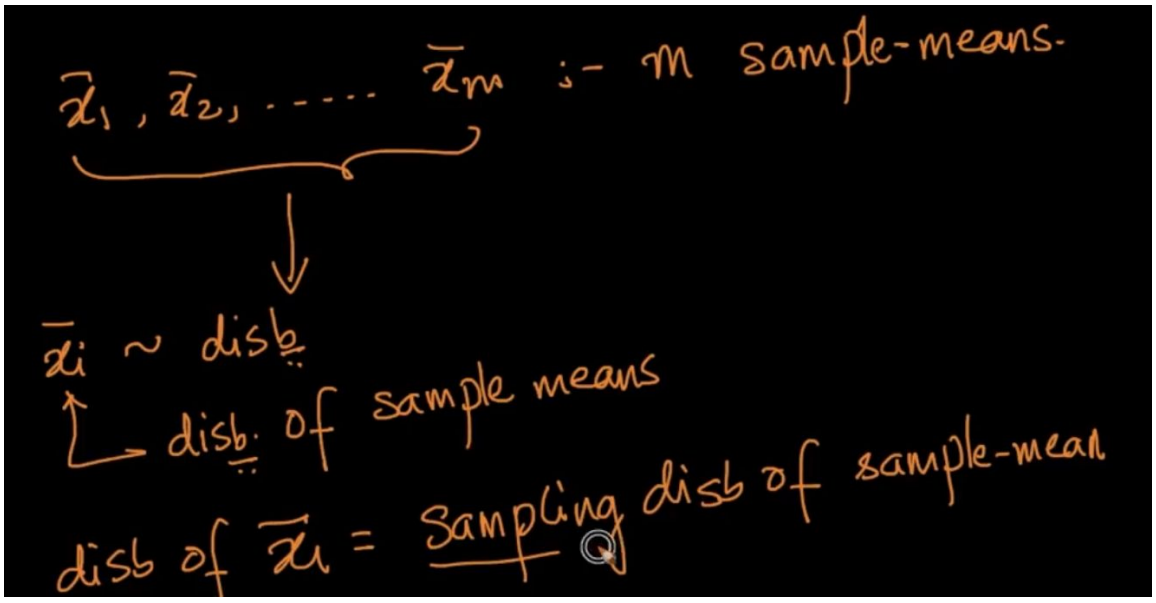
Kernel density estimate (KDE) with different bandwidths of a random sample of 100 points from a standard normal distribution. Grey: true density (standard normal). Red: KDE with h=0.05. Black: KDE with h=0.337. Green: KDE with h=2.

CENTRAL LIMIT THEOREM

Sampling Distribution

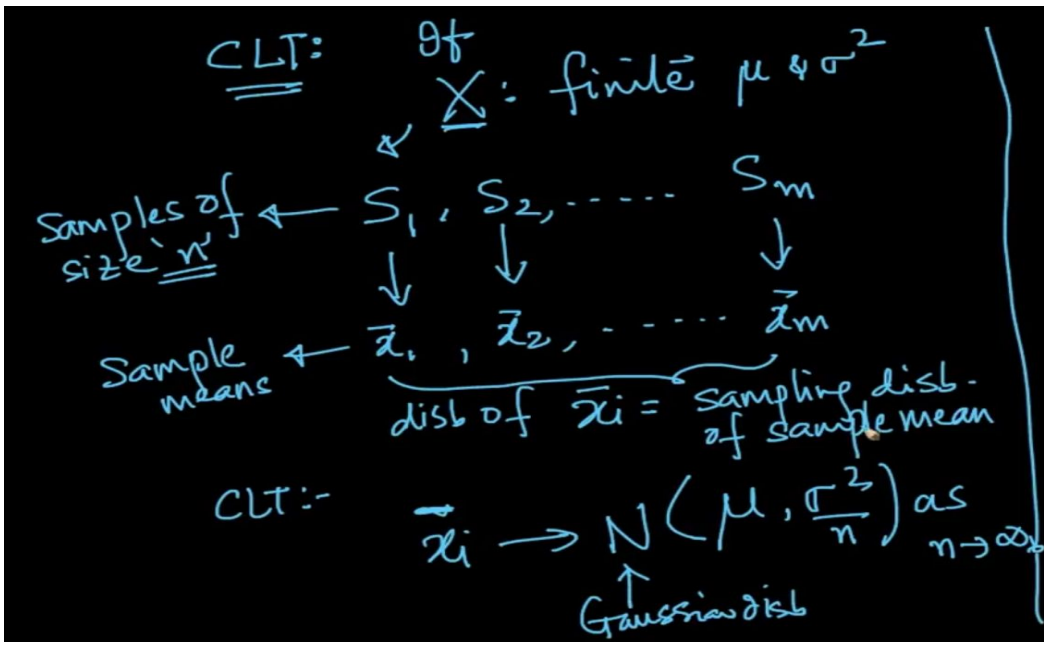


Suppose we've a population of X (disb. Of incomes) and we take out 'm' samples and their sample means $\bar{x}$ of each sample S.



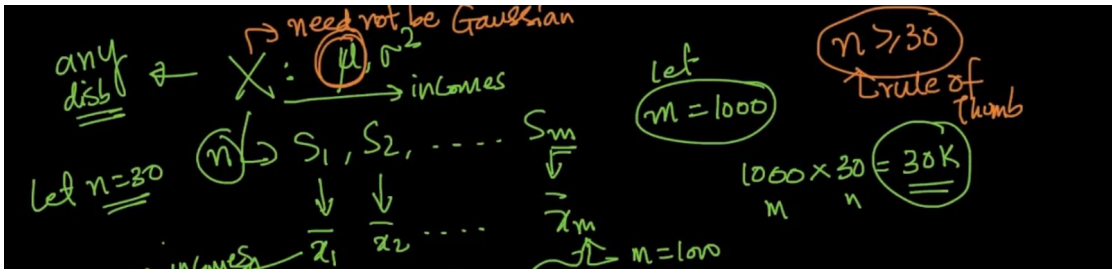
If we get sample means of each sample the distb of sample means is $\bar{x}_i$ and disb of $\bar{x}$ is called sampling disb of sample mean

CLT



CLT says that $\overline{xi}$ is Gaussian distribution (μ, $\frac{\sigma^2}{n}$) if X (population) : finite μ & σ^2

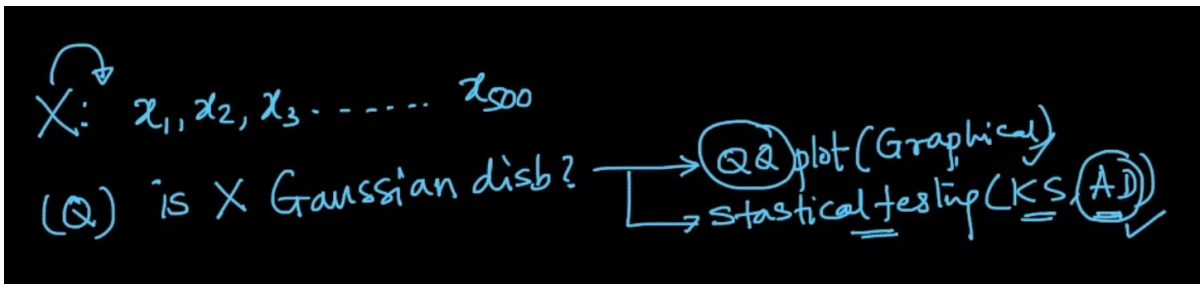
Explanation



We take out 30 samples from population X: μ, σ^2 and get the means and plot them we get a Gaussian disb with mean $\approx$ *population mean* μ and variance $\frac{\sigma^2}{n}$. Which can help us estimate the mean and variance of whole population X .Here ,with 30k we estimated mean of 7B population

QQ PLOT

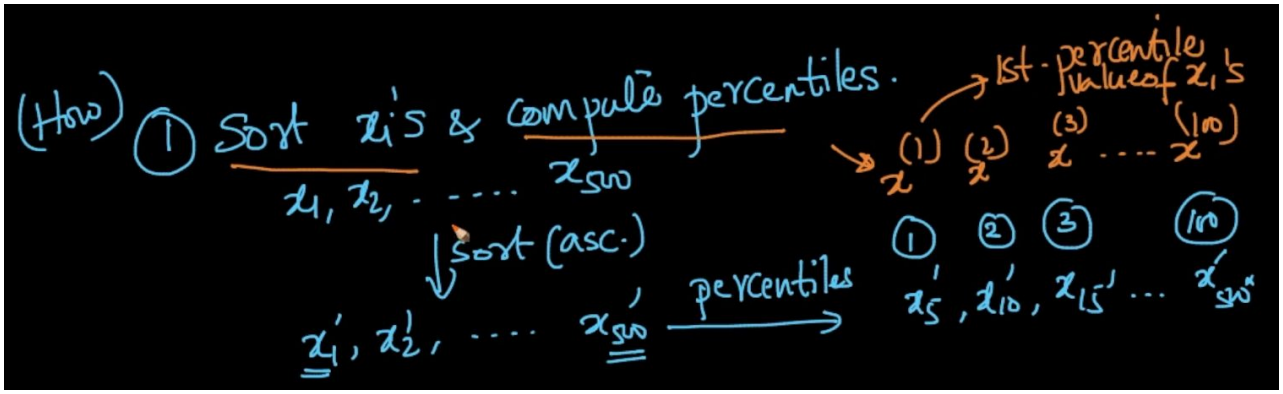
To test if random variable is normally distribution or not



There are two methods mentioned above. We'll see QQ plots

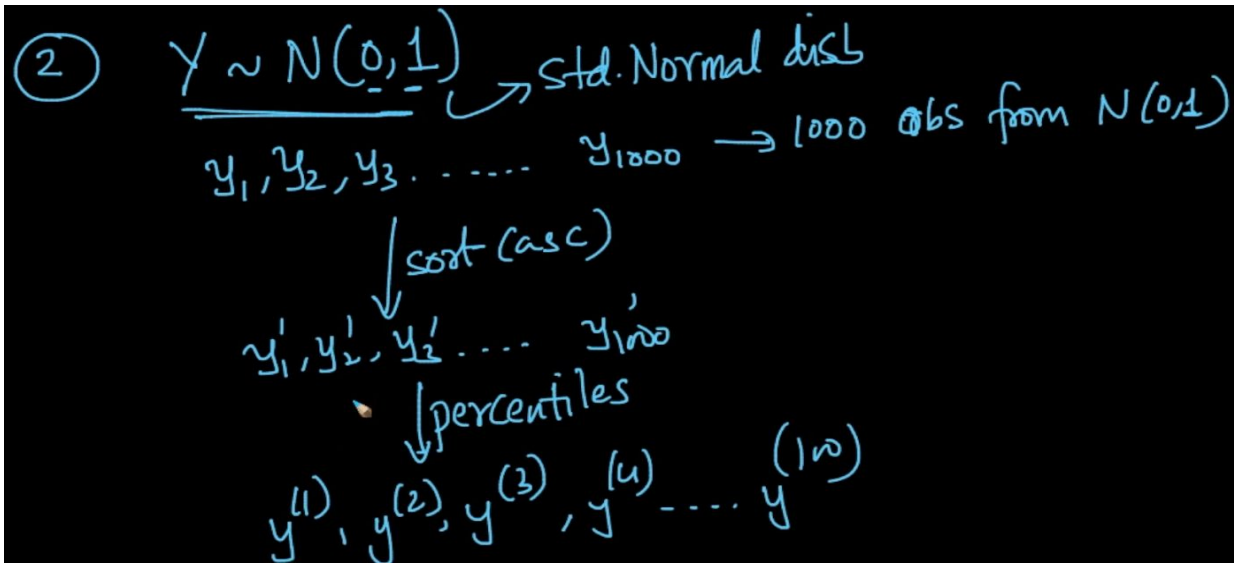
Steps of making QQ plots

Step 1



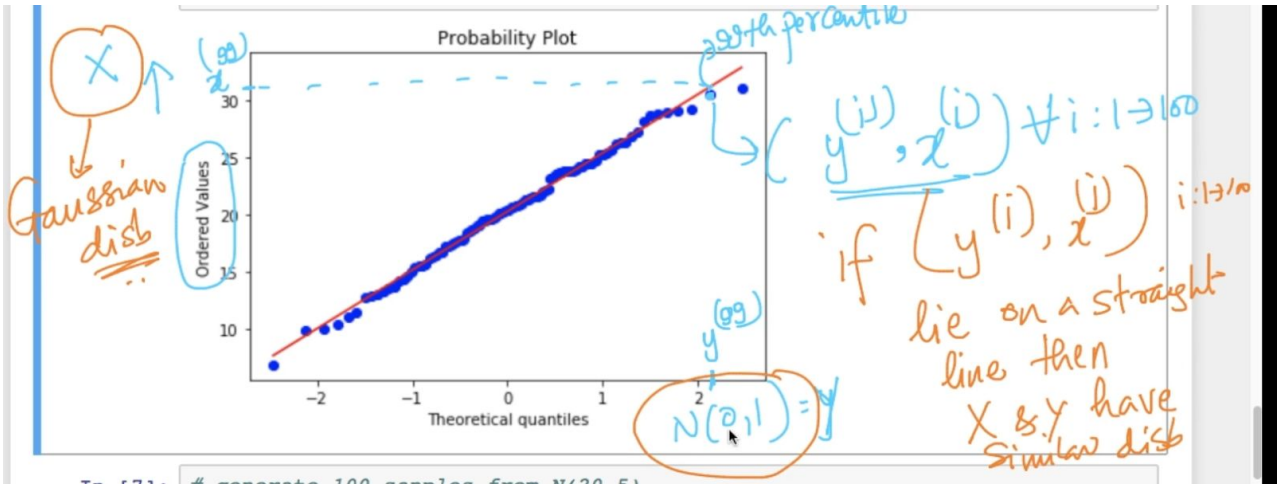
Sort x and compute their percentiles

Step 2



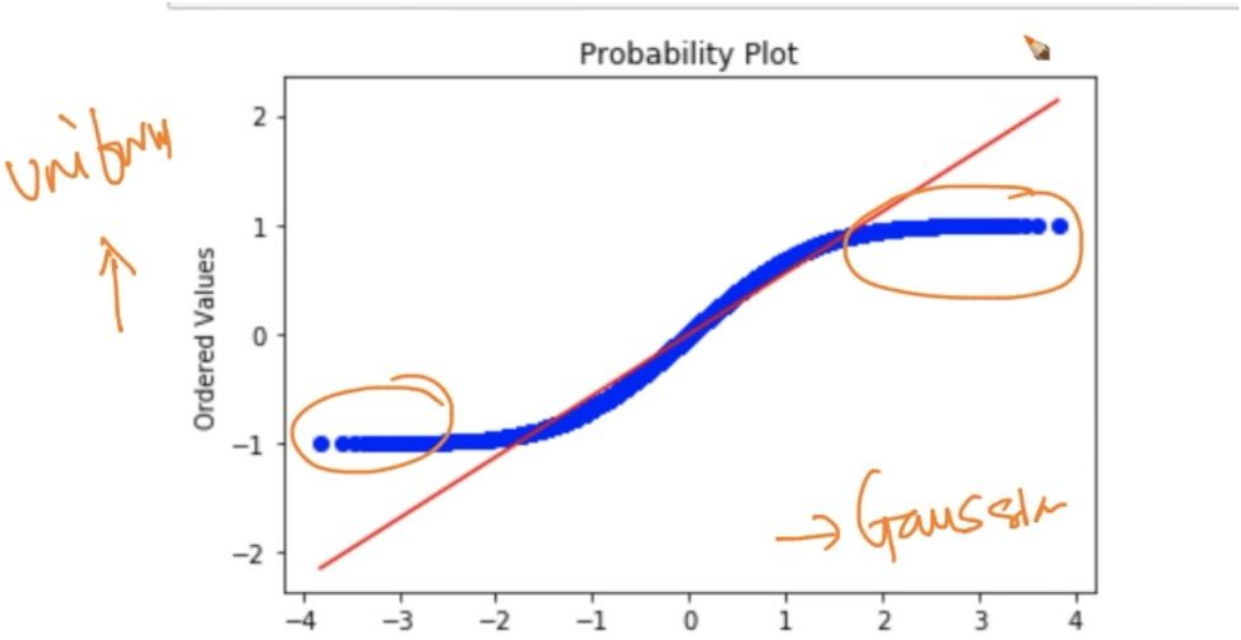
Create a r.v. Y with std normal disb as shown and calculate percentiles as shown above

Step 3 (Plot)



If (y^i, x^i) lie on a straight line then X and Y have similar distribution

Note (Limitation)- If sample size is small (let's say 50) then they are hard to interpret so it should be large so that they fit well (10000).

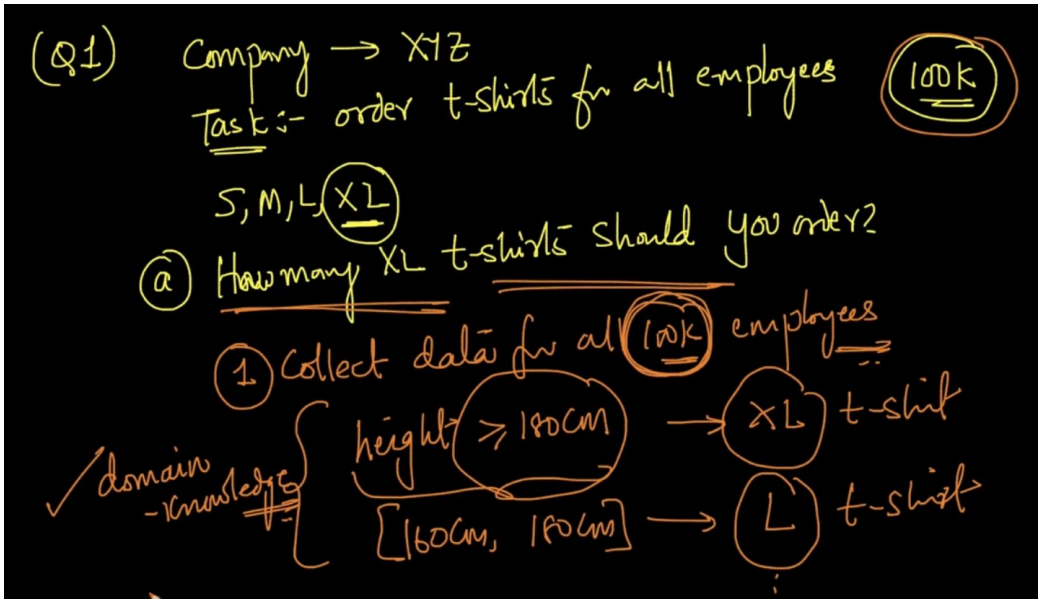


From the above fig. it can be seen that line is not fitting because our X (sample size : 50000) is uniform distribution and not Gaussian as more and more pts are furthering away

So , using QQ plots we can know that the two random variables share the same distribution or not.

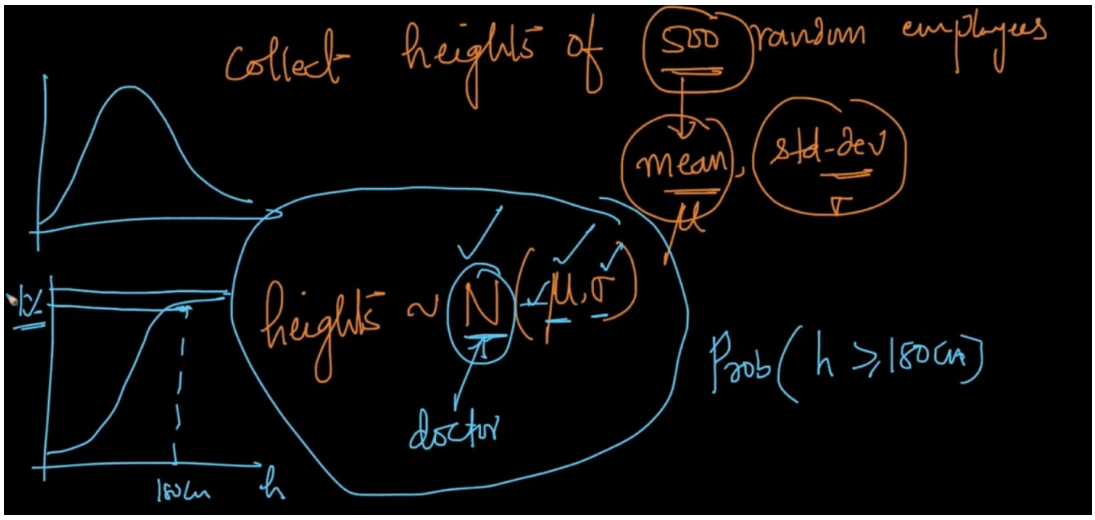
How distributions are used ?

Example



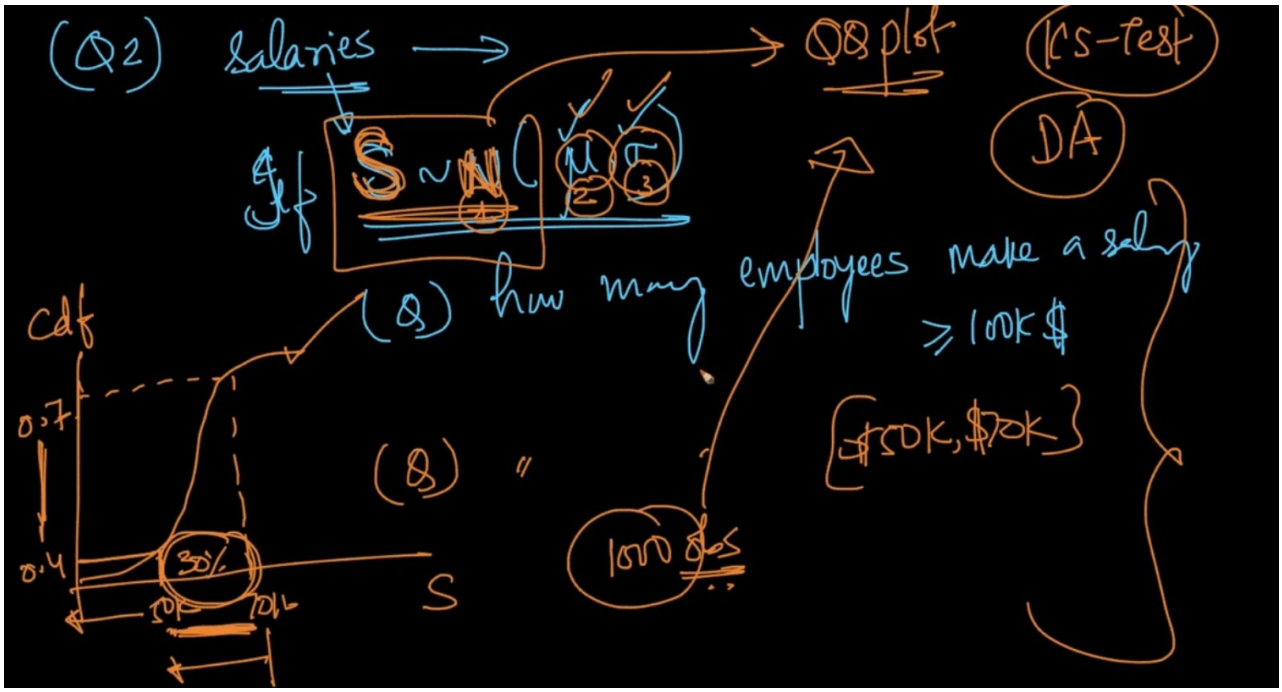
Collecting data for all 100k employees is costly and time consuming

We know that people whose ht > 180 cm -> XL



If we collect the heights of 500 random emp and our heights are Normally distribution then we can get the mean and variance and from that we can plot CDF and 1% > 180cm of people

Example 2

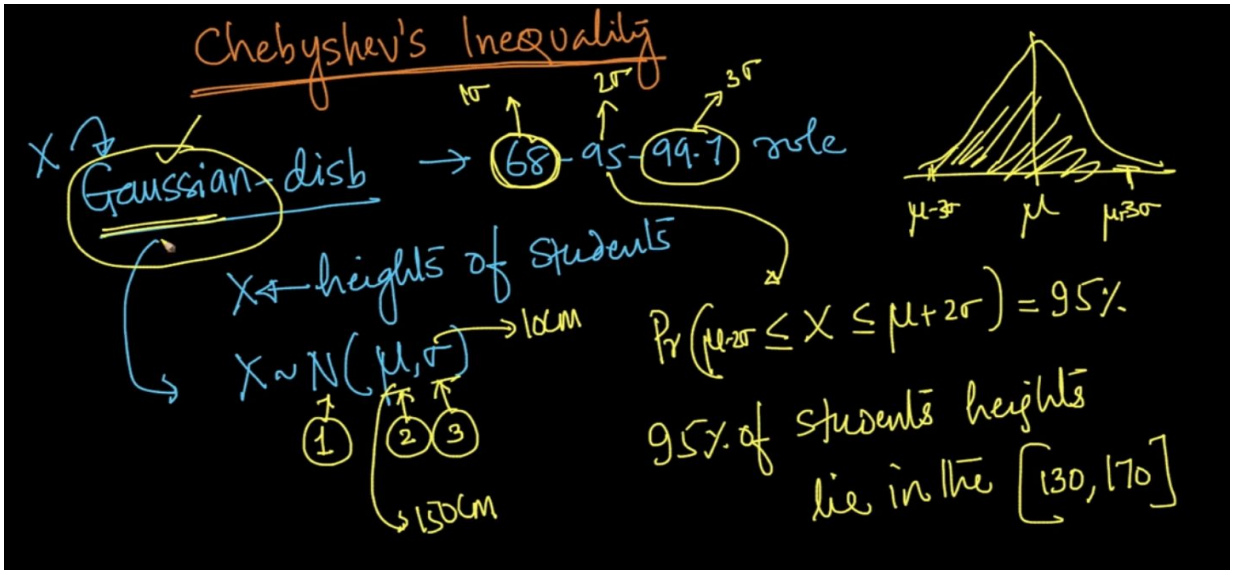


We can get idea of the above questions if our Salaries S is normally distribution by using a CDF graph

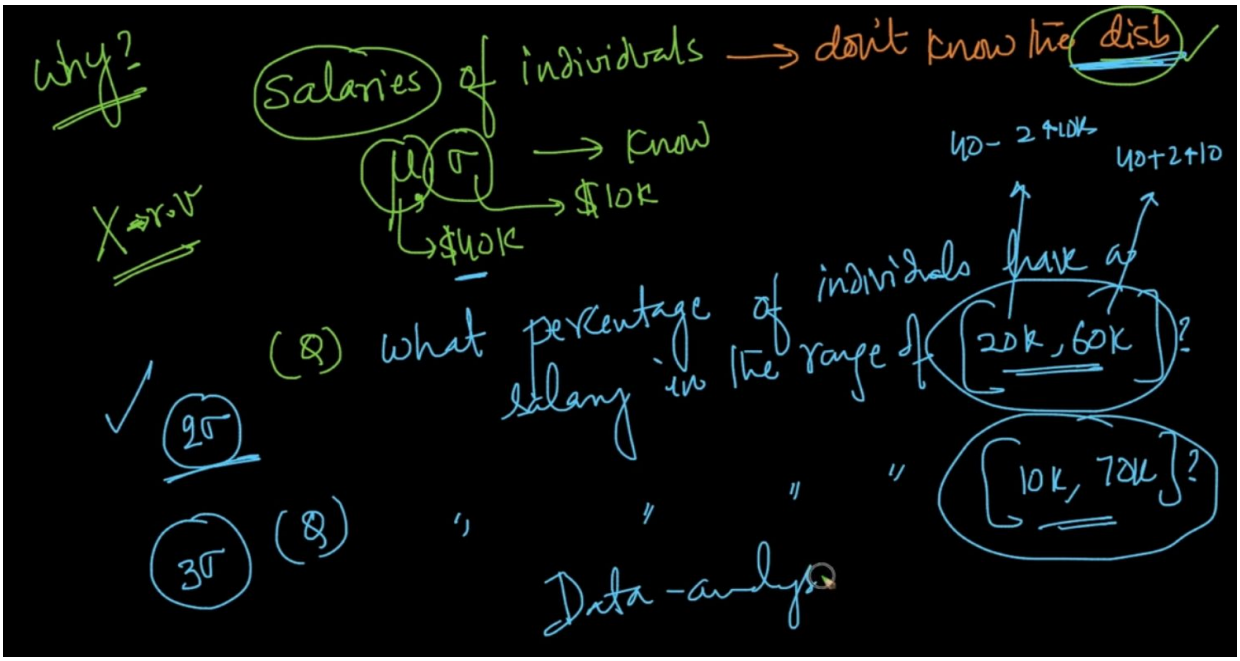
How to know if it's a Gaussian Distribution?

- By plotting a QQ plot as we know. There are some tests like KS- testing which will be studied later

CHEBYSHEV'S INEQUALITY



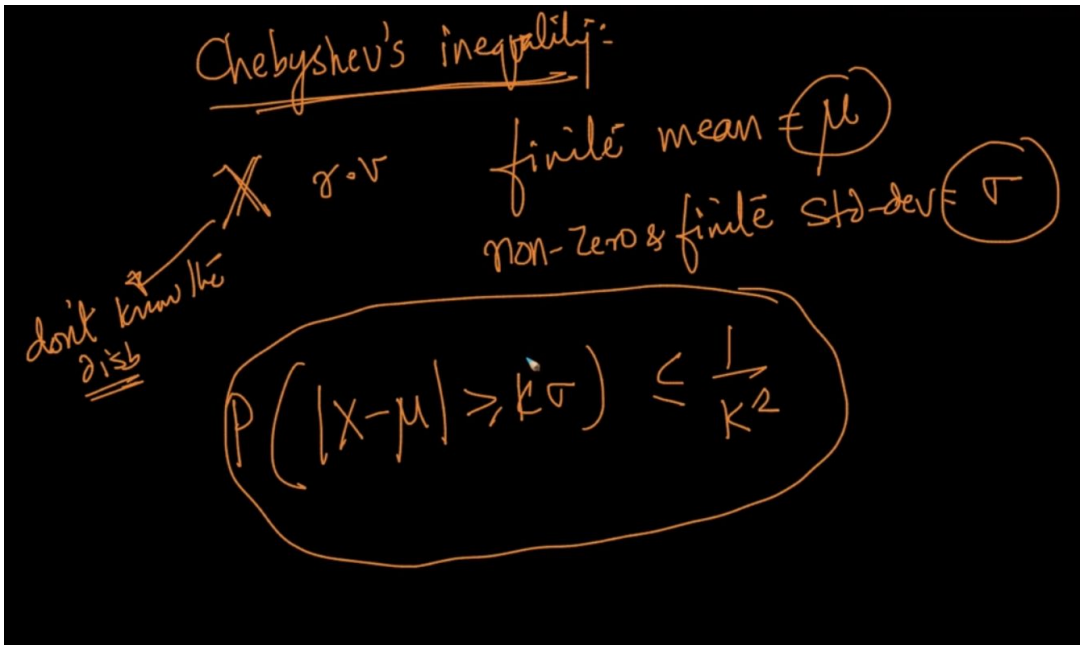
We got all the answers from the 68-95-99.7 rule if we know that our r.v is Gaussian distribution but what if it isn't?



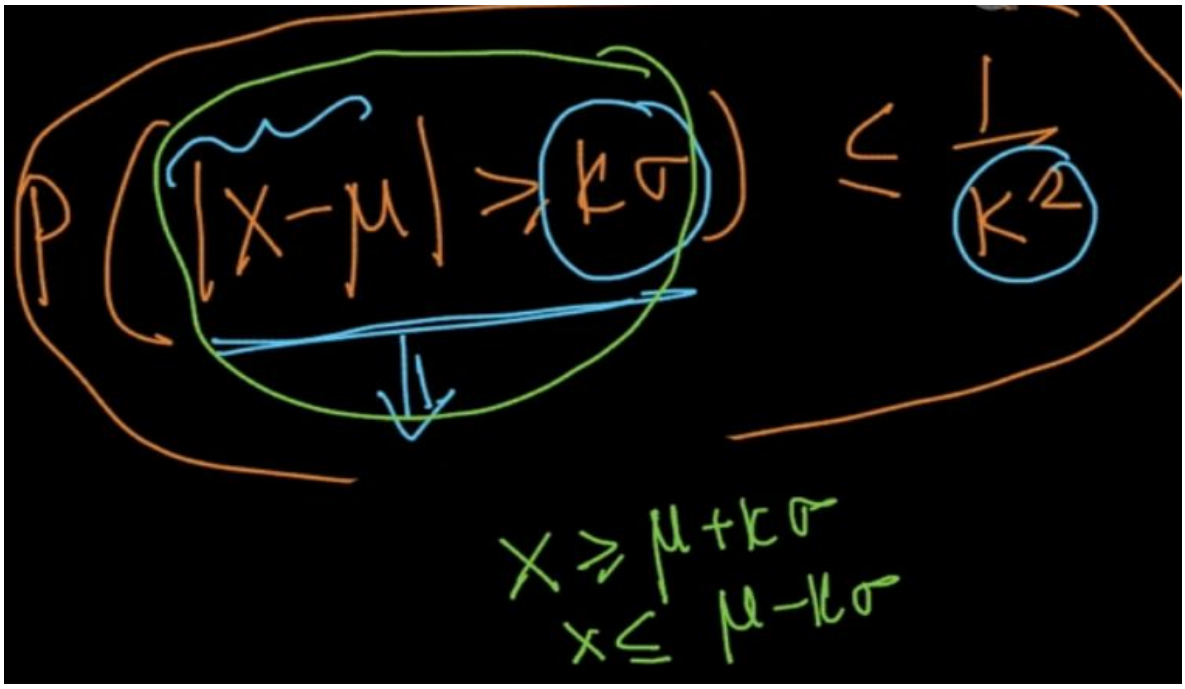
In the above ex. , if we dont know the disb but we know the μ *and* σ then what?

If it'd been Gaussian then the answer woold have been pretty simple that 95% of our data lies within 2 σ *range* [20*k* − 60*k*] but here we don't know the disb so what now?

The answer is Chebyshev's inequality the equation is shared below



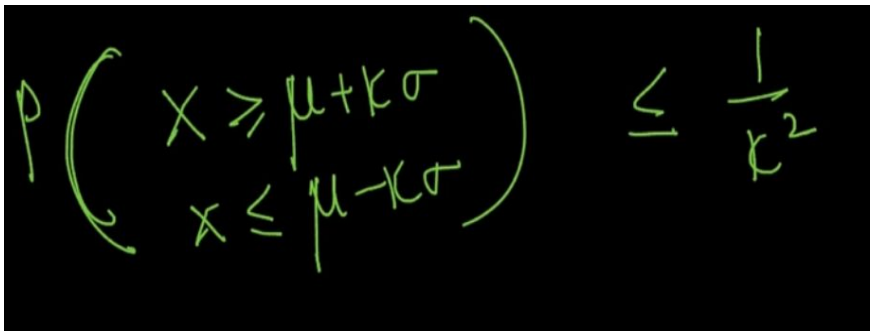
Let's break down the equation

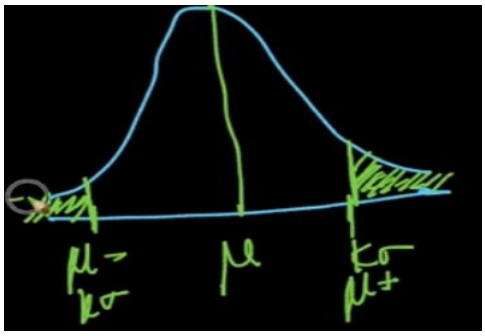


Property of mod reminder to understand the above

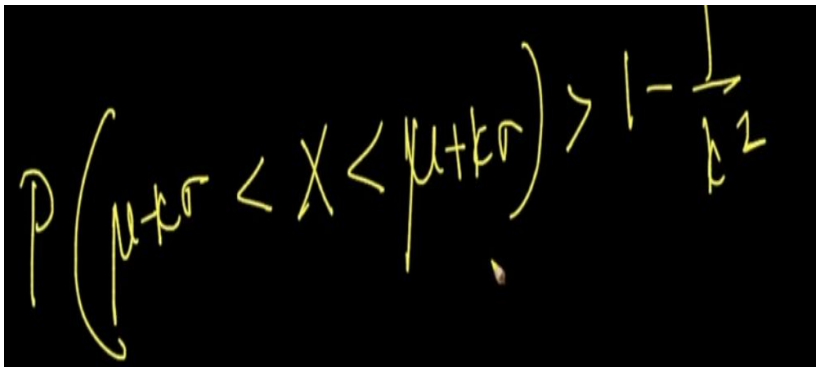
If | X | > y (any_no) then X > y or X < y

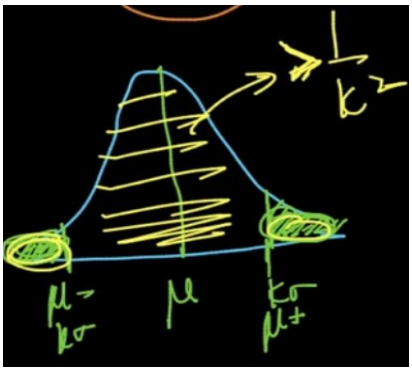
For ex. If |X| = 20 and y =15 then X > y or X < y since X can be +20 or -20



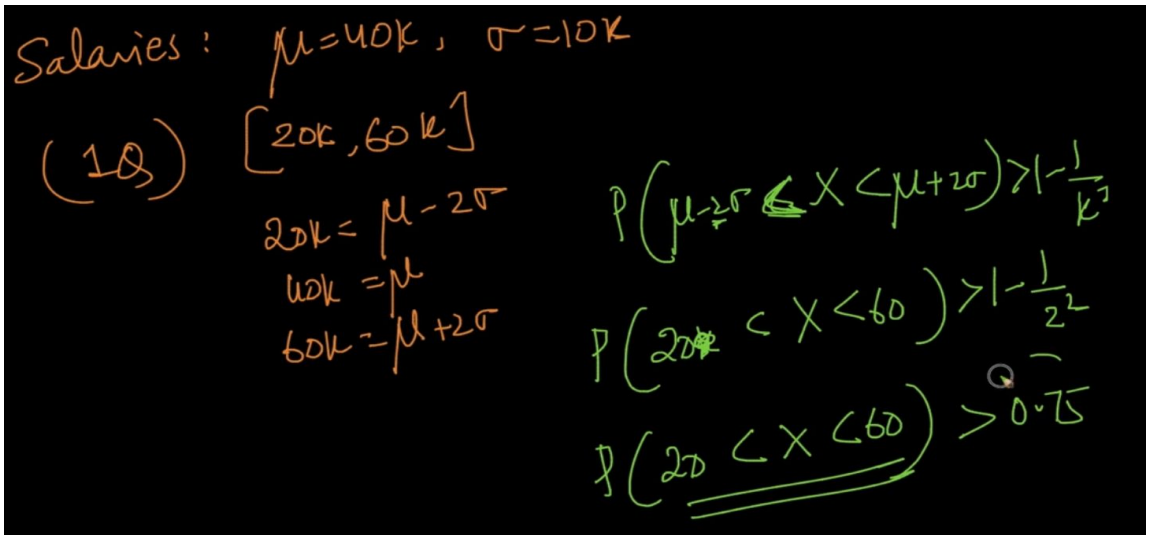


The green region is what this Probability is implying (example below will clear all)



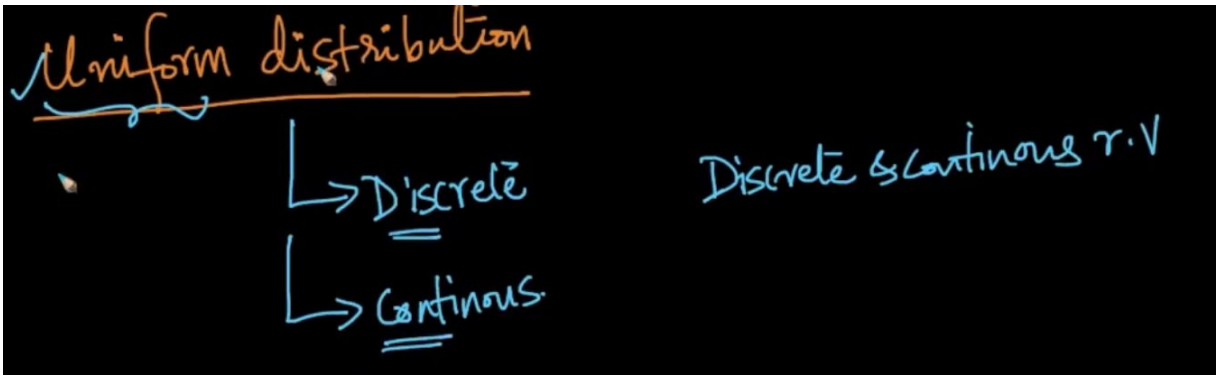


The yellow region is greater than $\frac{1}{k^2}$ because the probability of the green region is less than $\frac{1}{k^2}$

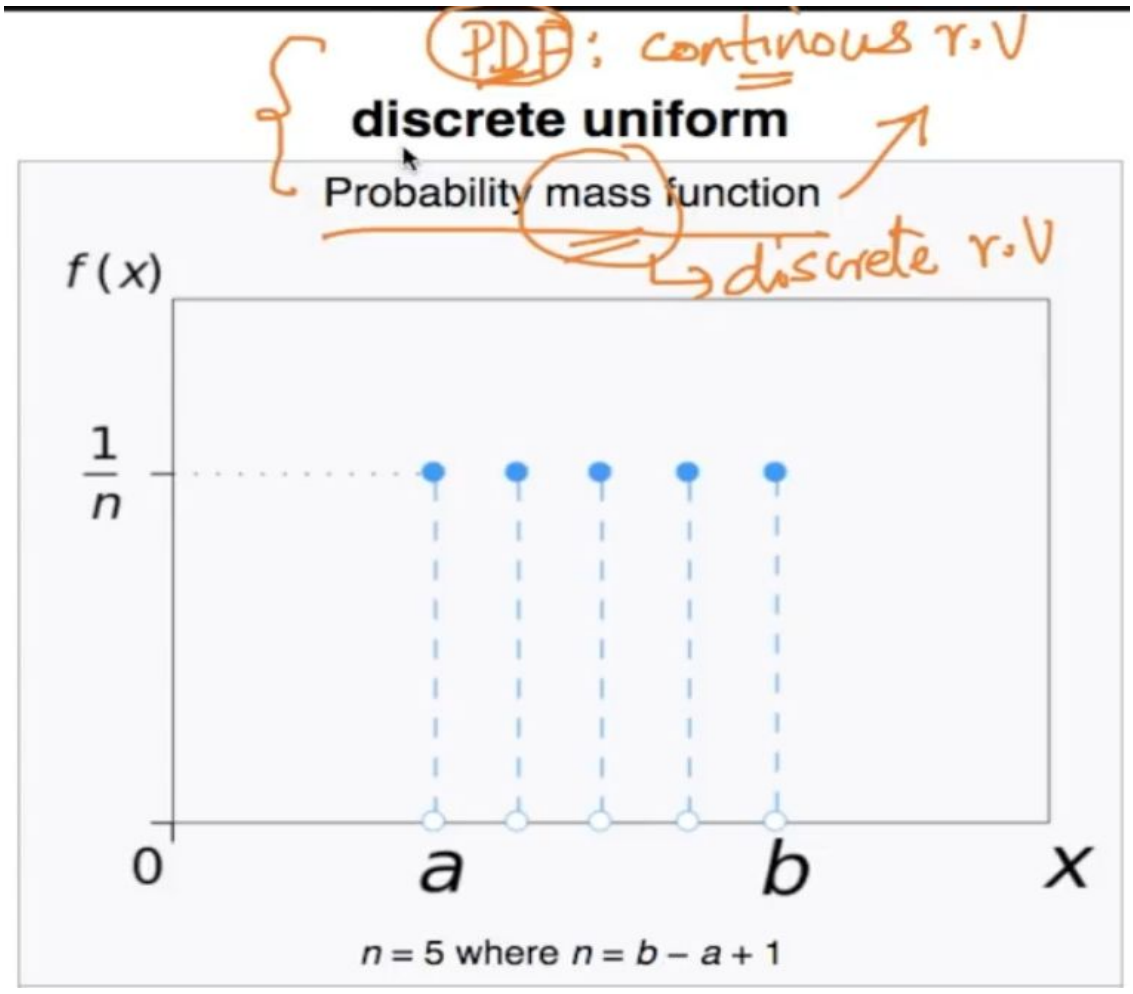


We dont know the disb here but we know our probability . If it'd been Gaussian then it would be 95% but here without even knowing disb we know that 75% lie between (20k, 60k)

UNIFORM DISTRIBUTION

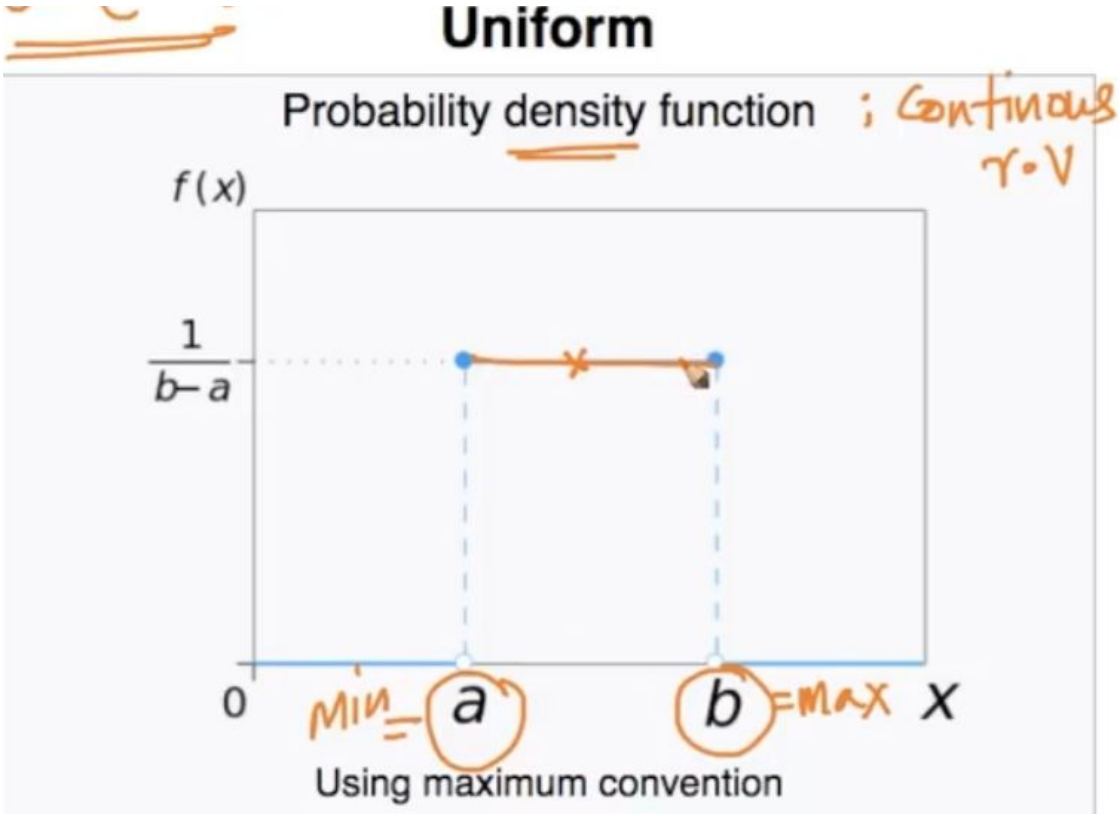


Two types



The height of the PDF/PMF gives probability of that r.v. As we can see that in the distribution every r.v has the same probability / equiprobable. Where parameters are ‘a’ and ‘b’

CONTINUOUS UNIFORM DISTRIBUTION



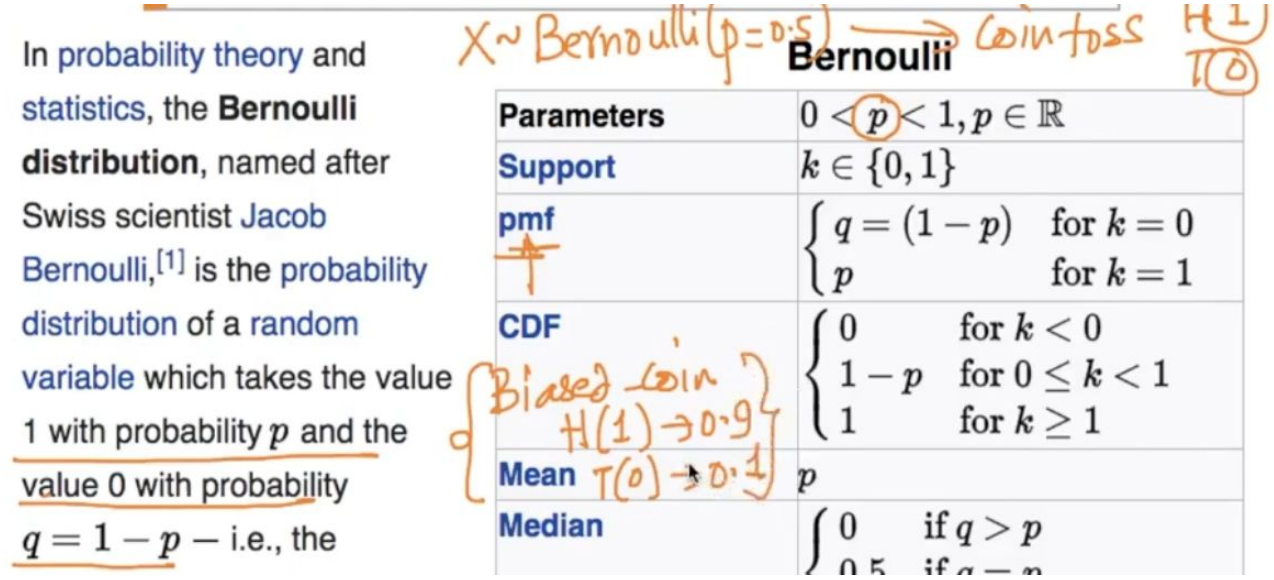
The probability of any r.v between a and b is same which is $\frac{1}{b-a}$ because the area under the curve is 1 so height will be $\frac{1}{b-a}$ since width is (b-a) (area = breadth * height)

Uniform Distributiion (continuous)

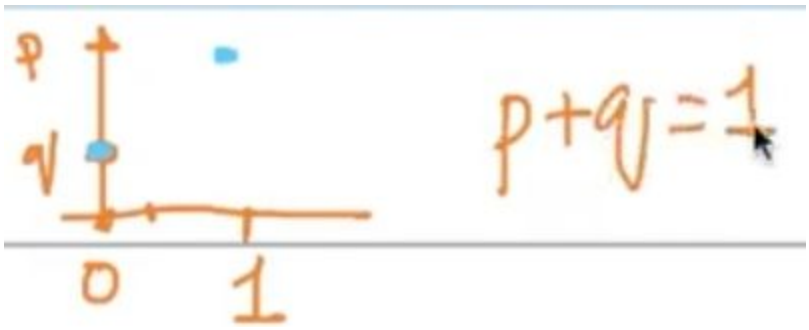
Uniform distribution (Discrete)

BINOMIAL AND BERNOULLI DISTRIBUTION

BERNOULLI DISTRIBUTION



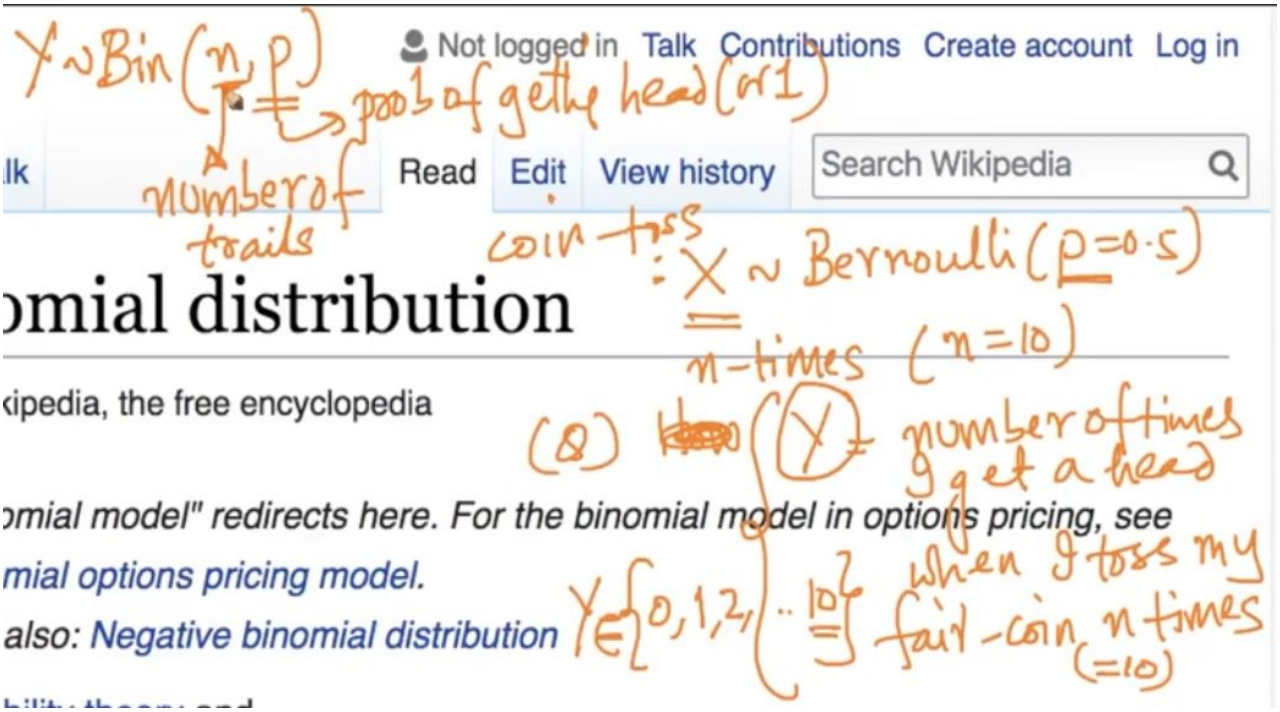
Bernoulli can only have two outcomes with p and q equalling 1 in total.



Pmf of distribution random variable. Just remember that the sum of it is 1.

[Bernoulli distribution - Wikipedia]

BINOMIALDISTRIBUTION



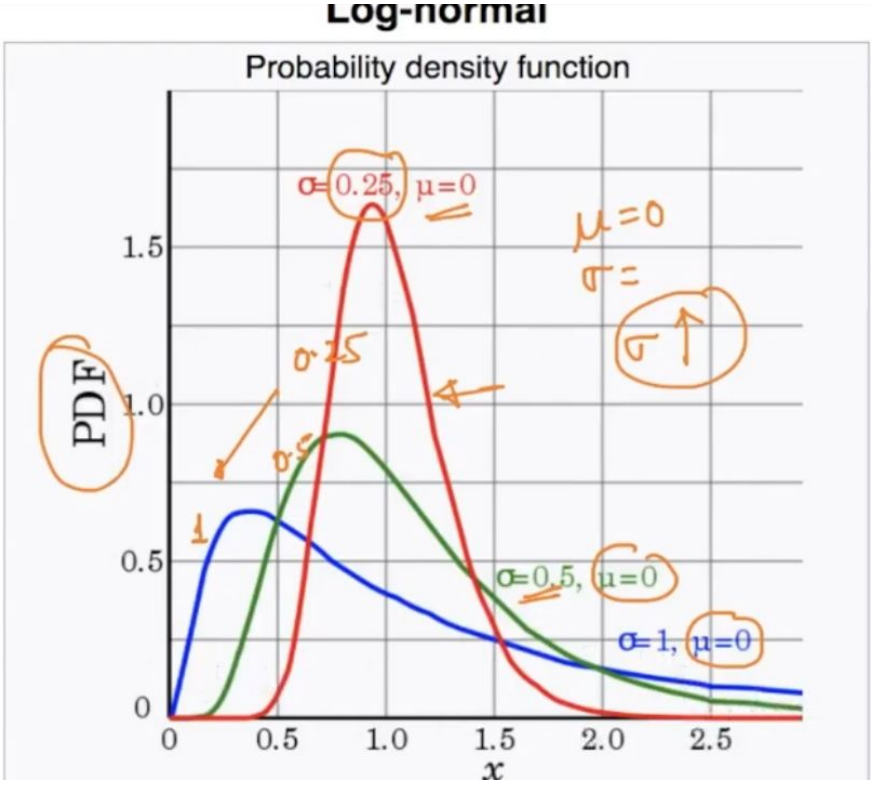
Let's take the example of Coin toss where p = 0.5 (probability). We can see the question above for Y. It can be seen that the values can be between Y ε {0,1,2,3….10} for the above question

Notation	$B(n, p)$
Parameters	$n \in \mathbb{N}_0$ — number of trials
	$p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$

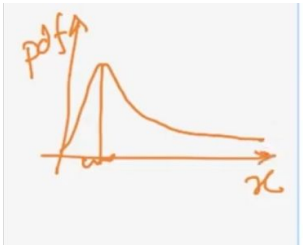
Pmf - It says the Probability of Y = k is the above. Binomial distribution - Wikipedia

LOG DISTRIBUTION

A random variable X is said to be log-normally distributed if log(X) is normally distributed.



As it can be seen it is largely skewed at the right side. This is log-normally distribution

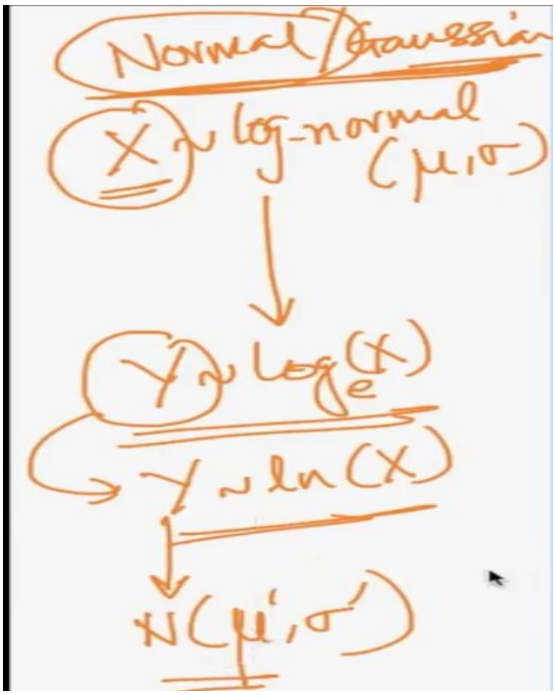


Examples include the following:

- Human behaviors
  - The length of comments posted in Internet discussion forums follows a log-normal distribution.[18]
  - The users' dwell time on the online articles (jokes, news etc.) follows a log-normal distribution.[19]

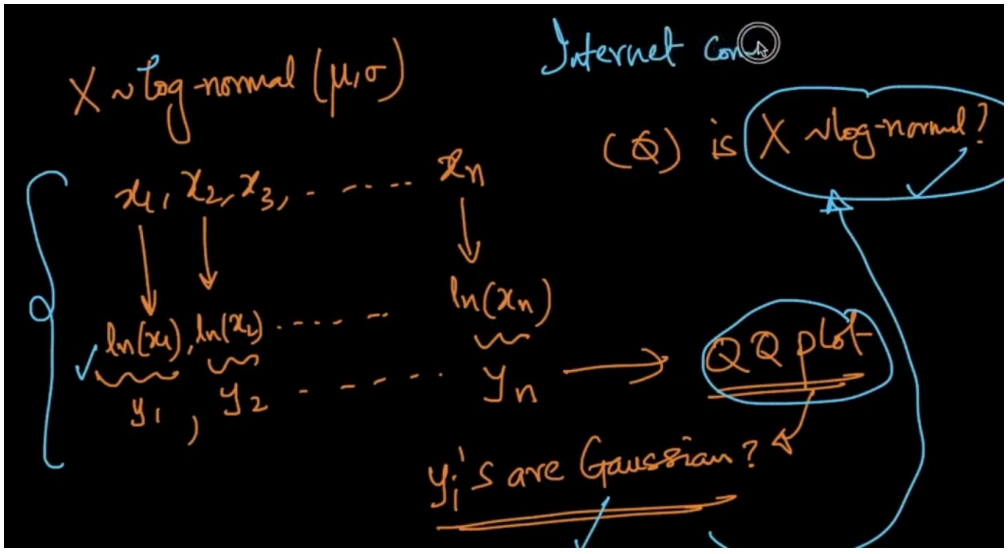
reddit, Quora

This is log-distribution of comments in a post. Most people comment short but some people comment long.



It's been prevelant that if X is log normal (μ, σ) then convert it into log ->(log(X)) after that we can convert this into Normally distribution

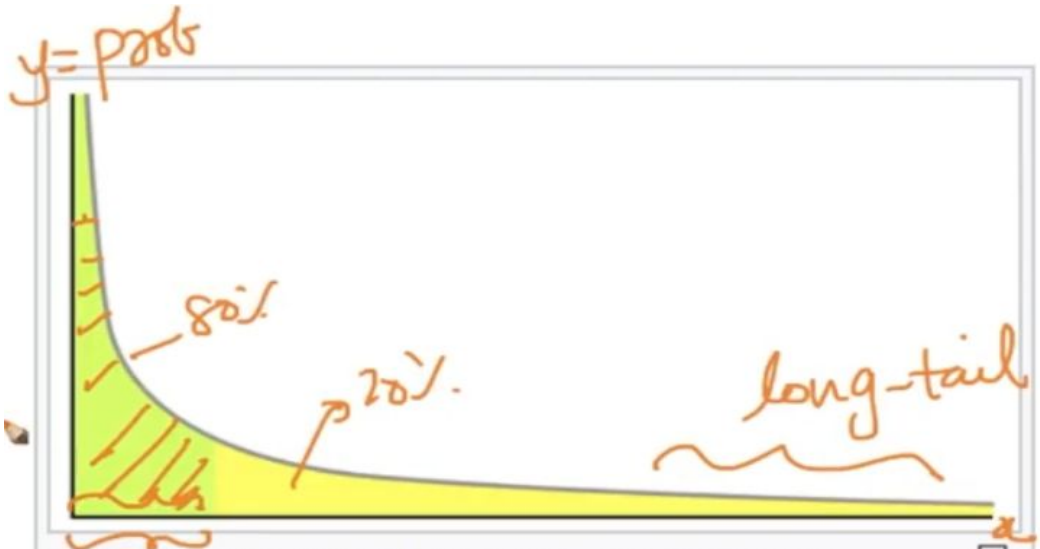
How to know X is log normal distribution?



We take the log of values X and then plot a QQ plot of Y if it's Gaussian then X is log normal

NOTE - In the Internet log normal happens in a lot of times but Normal distribution occurs only a few times . [Log-normal distribution - Wikipedia]

POWER LAW DISTRIBUTION



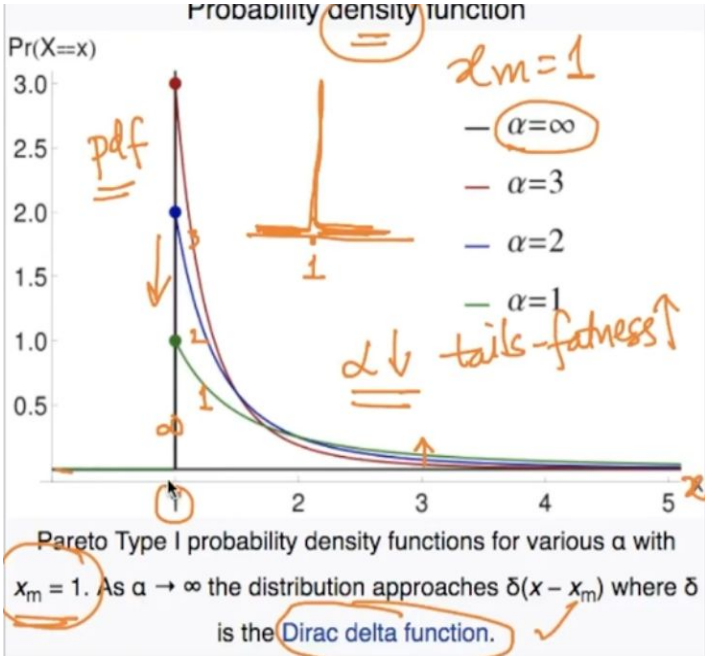
An example power-law graph, being used to demonstrate ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

In the Power-law distribution , 80% of the mass/density lie in 20%. Whenever a distribution follows Power-law it follows Pareto distribution

Power law - Wikipedia

Pareto distribution - Wikipedia

PARETO DISTRIBUTION



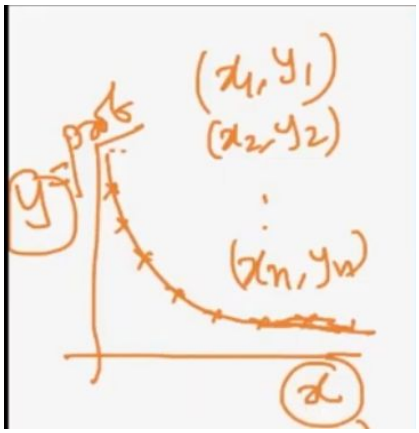
Parameters :- $x_m > 0$ & $\alpha > 0$ take these parameters as mean(x_m) and variance (σ)

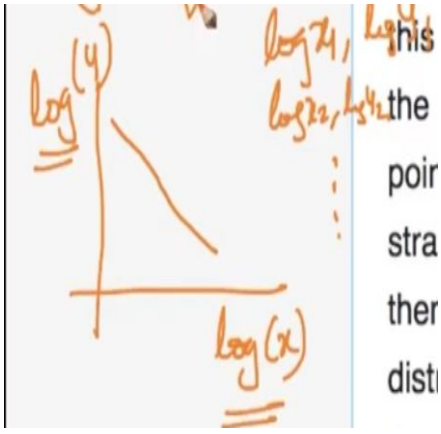
In the above image $x_m = 1$ and as $\alpha \downarrow$ $tail-fatness \uparrow$

Note - If $\alpha = \infty$ then it's just 1 peak as drawn above . It is called Dirac- delta function

Applications - 1) The values of oil reserves in oil fields (a few large fields, many small fields)

How to check if our c.r.v is Power law?





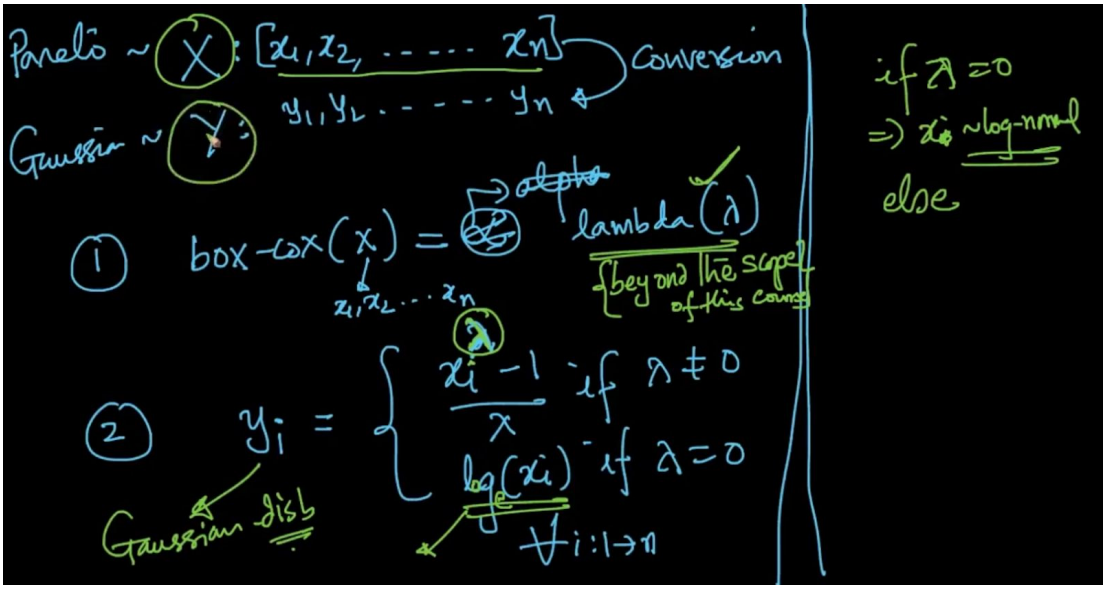
If x (input) and y(Probability) then take out the **log** of both x & y and if straight line then Power Law .

QQ plots can also be used to check if it follows PAreto Distribution

BOX COX TRANSFORM

We've seen that if X is log-normal then it can be converted to Y (Gaussian) by transforming X to log but how will we convert if X is a power-law/ pareto distribution to Y (Gaussian)

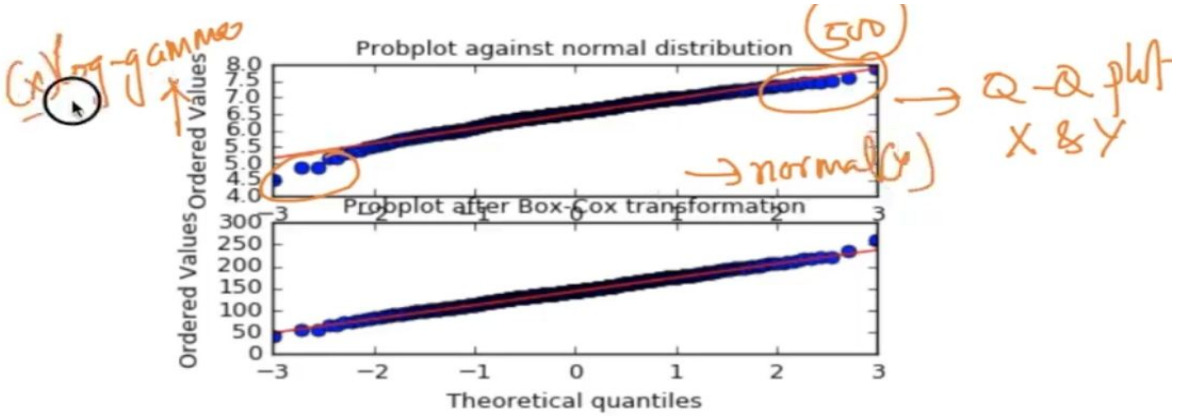
How will we transform ?



1) We will try to run a function called box-cox to our r.v X and get lambda (λ)

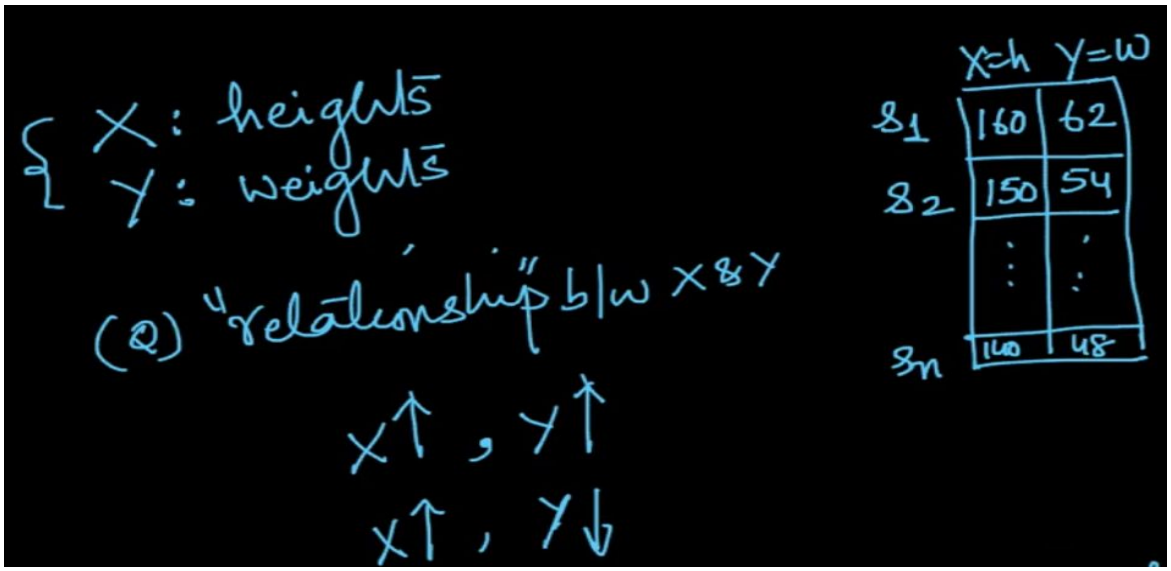
2) We get the above y (Gaussian) with the formula and then our power-law distribution will get converted to Gaussian distribution.

Note - if λ = 0 then x -> log normal



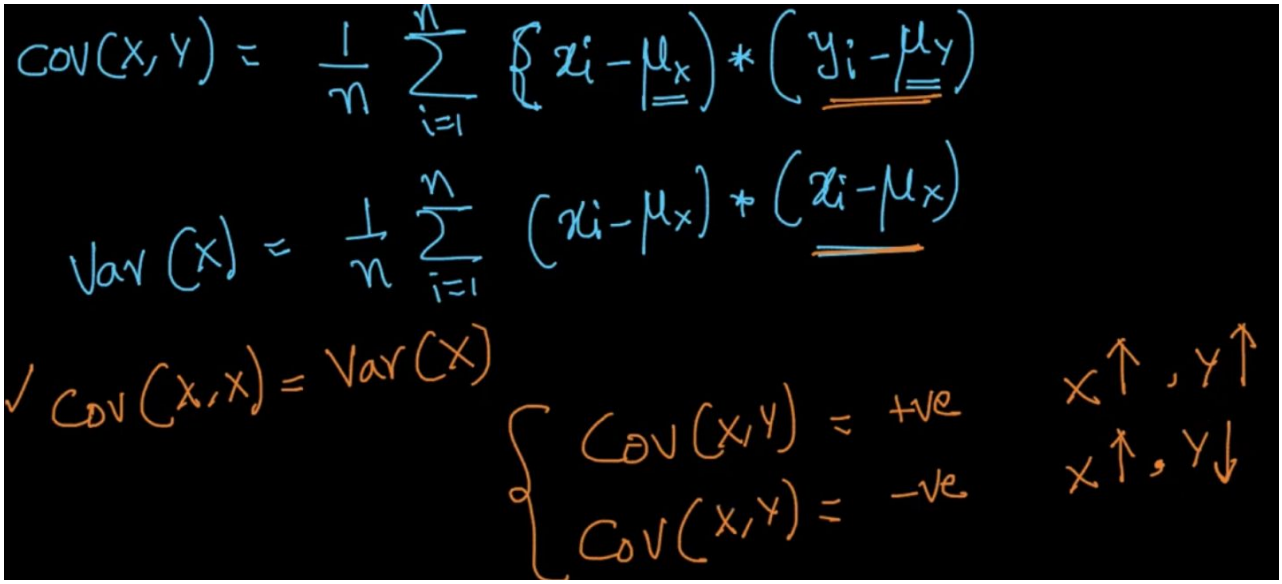
With QQ plots we can check if it's normal disb or not. X was log gamma and Y normally distributed but after applying box-cox to X it is converted to Normal.

CO VARIANCE



Suppose we've X : heights, Y: Weights then if we need to find the relation between them like if X↑ then Y↑ or X↑ *then Y* ↓ the answer to this is some correlations like Co-variance, PEarson corelation coefficient or Spearman co-relation coefficient

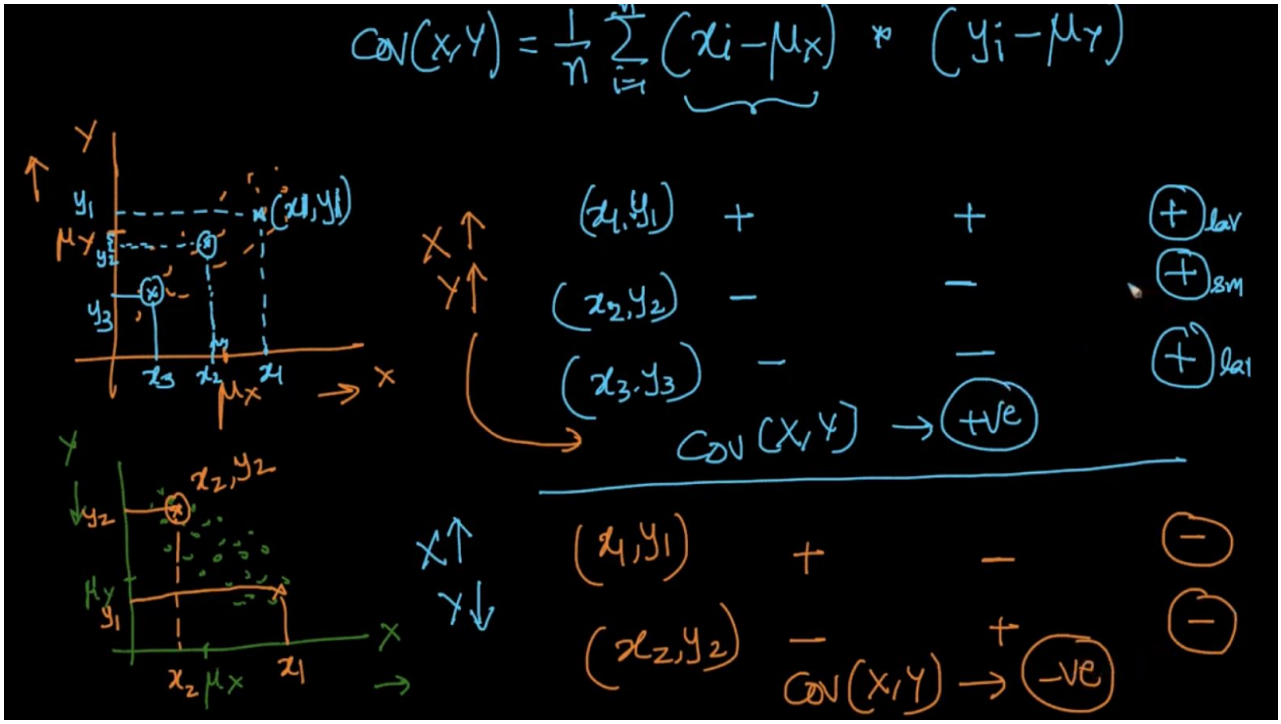
CO VARIANCE



Looking at the formula it can be clearly seen that variance(X) is covariance of X with itself

By getting the covariance if Cov(X, Y) = +ve then X ↑ , Y ↑

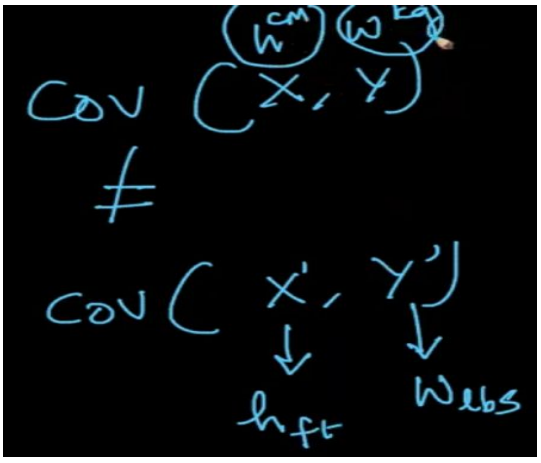
And if Cov(X, Y) = -ve then X ↑ , Y ↓



In the 1st img ,if ($x_i - \mu_x$) and ($y_i - \mu_y$) are both positive then Cov(X, Y) = +ve then as

X ↑ *Y also* ↑

Problem with covariance

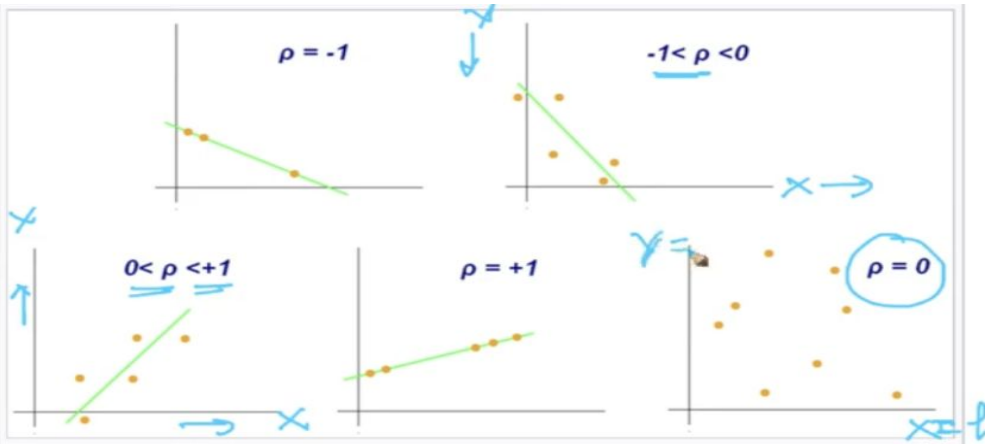


If we change from X (cm), Y(kg) to X(ft) , Y(lbs) then covariance wont be equal. Even when the dataset is the same as just it has calculated with different metric as mentioned.

PEARSON CORRELATION COEFFICIENT

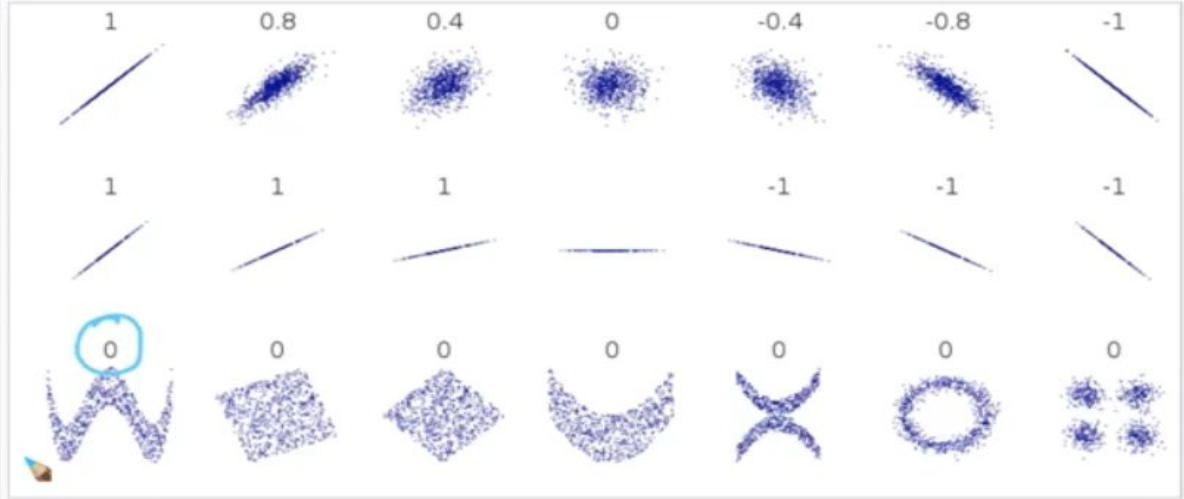
In PCC , Cov(X,Y) is divided by product of std.dev σ_x, σ_y . In covariance, .we got the idea that if both X and Y ↑ then Cov(X, Y) is +ve but we've no idea how +ve

PCC gives a good idea about that.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

The PCC (ρ) is always between -1 and 1. If ρ = 1 then it perfectly fits the line and X and Y are increasing. If -1 then decreasing. If it's not fitting then ρ is between -1 and 1 as seen above.

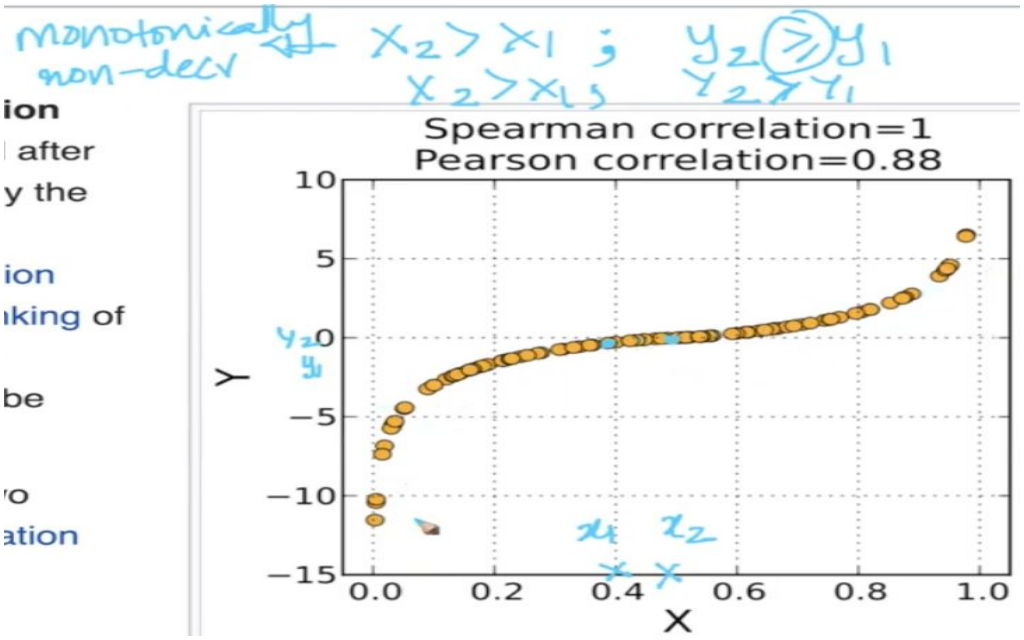


Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the non-linearity and

In the first row as line doesn't fit perfectly the ρ *decreases* .

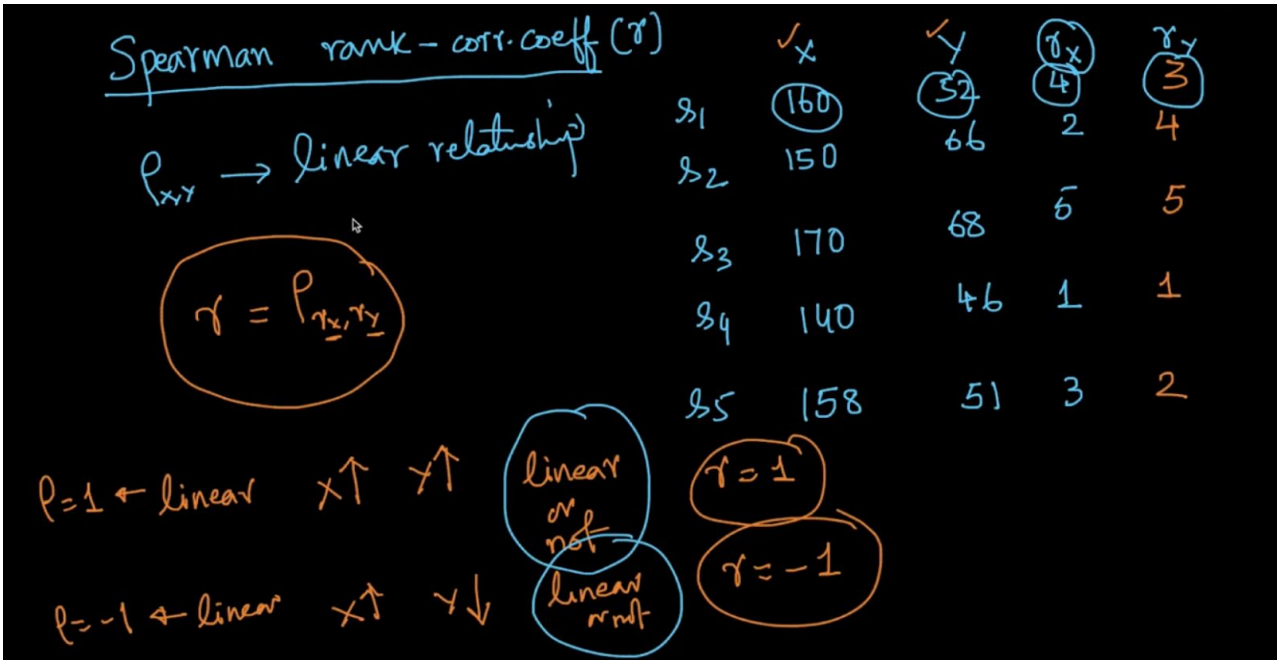
Second row: All are perfectly fitting. Note that it slope doesnt matter here.

Third row: PCC(ρ) doesn't work for Non-linear combos.



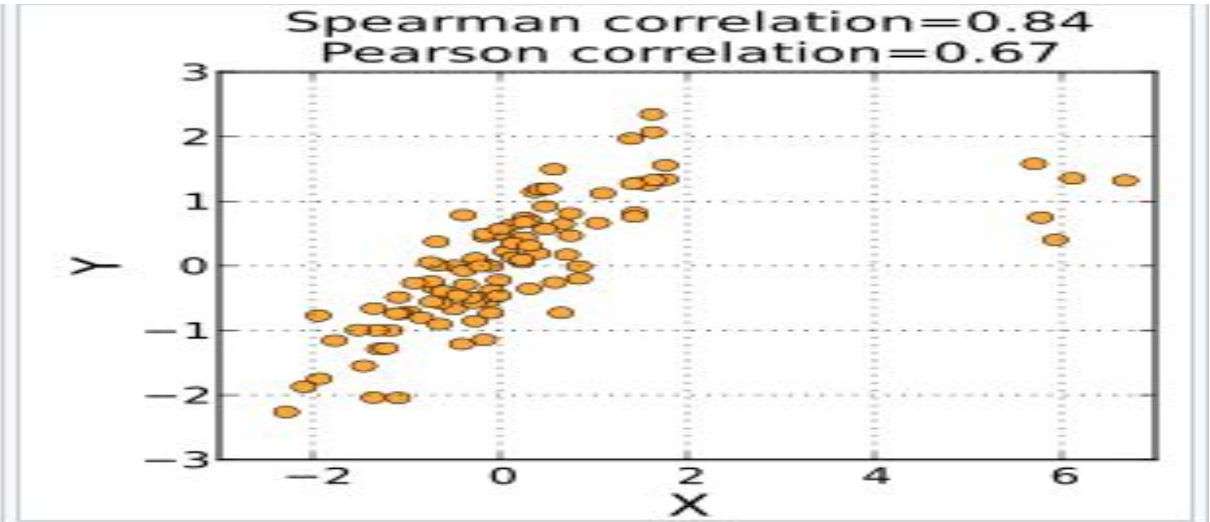
If X2>X1 ; Y2 ≥Y1 monotonically non decreasing. If X2 > X1; Y2 > Y1 monotonically increasing. In the image, Spearman co = 1 but Pearson co = 0.88 since not linear.

SPEARMAN RANK CORRELATION COEFFICIENT



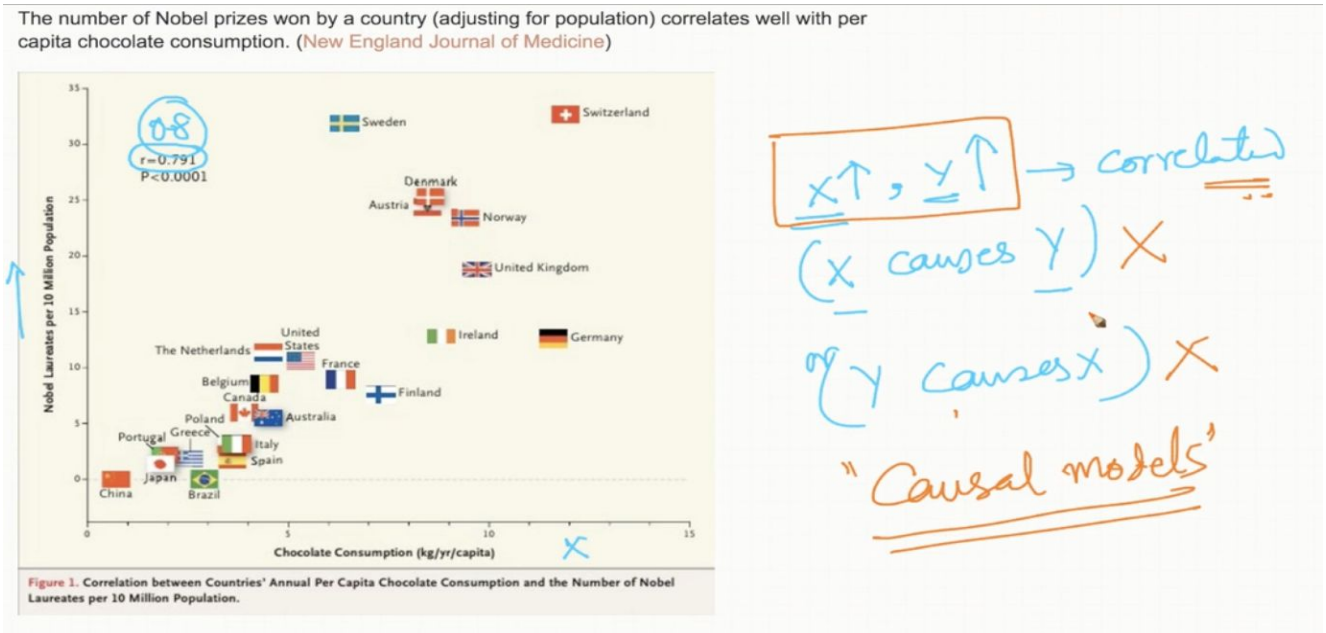
In Spearman rank, we dont us X and Y instead we sort them ascendingly and give them ranks as seen above and then r = ρ_{r_x,r_y} where r_x *and* r_y are ranks of X and Y. Spearman doesn't care about the figure following linear relationship like Pearson. It only cares about the ranks, so if both X and Y are ↑ then r = 1 whether it's linear or not. If X ↑, Y ↓ then r = -1

[Spearman's rank correlation coefficient - Wikipedia]



The Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples. That is because Spearman's rho limits the outlier to the value of its rank.

CORRELATION v/s CAUSATION



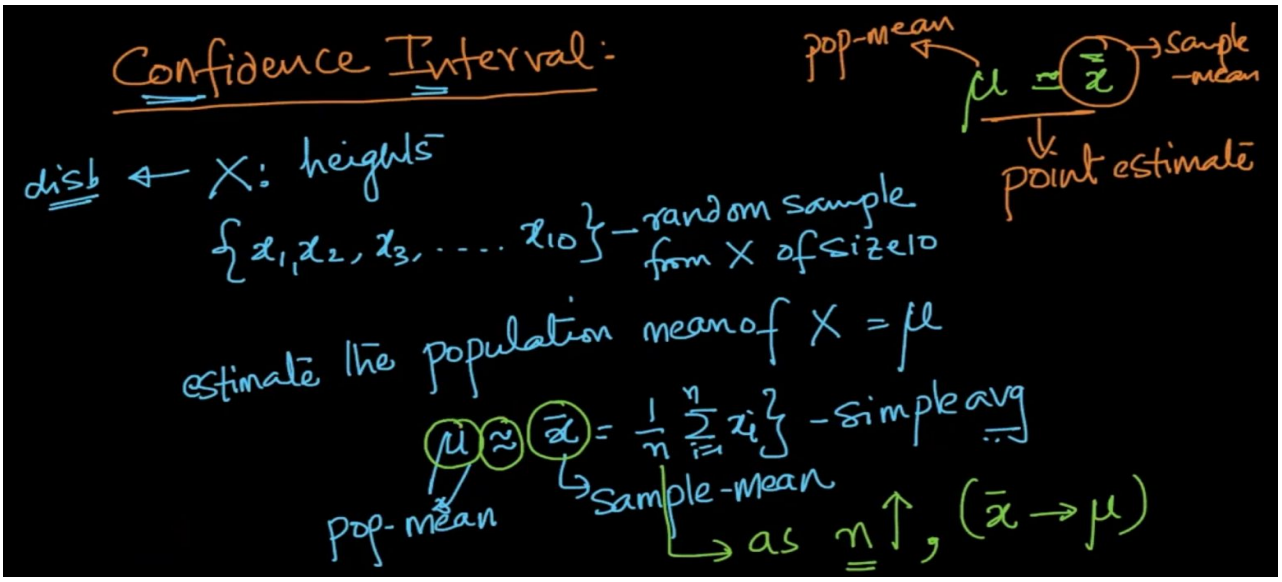
Data of Chocolate consumption(X) v/s Nobel Laureates per 10M population(Y). As X ↑; *Y* ↑ but that doesnt imply that X is causing Y or vice versa.

Causation is advanced statistics topics in which you learn about causal models which tells us what is causing what

How to use correlations?

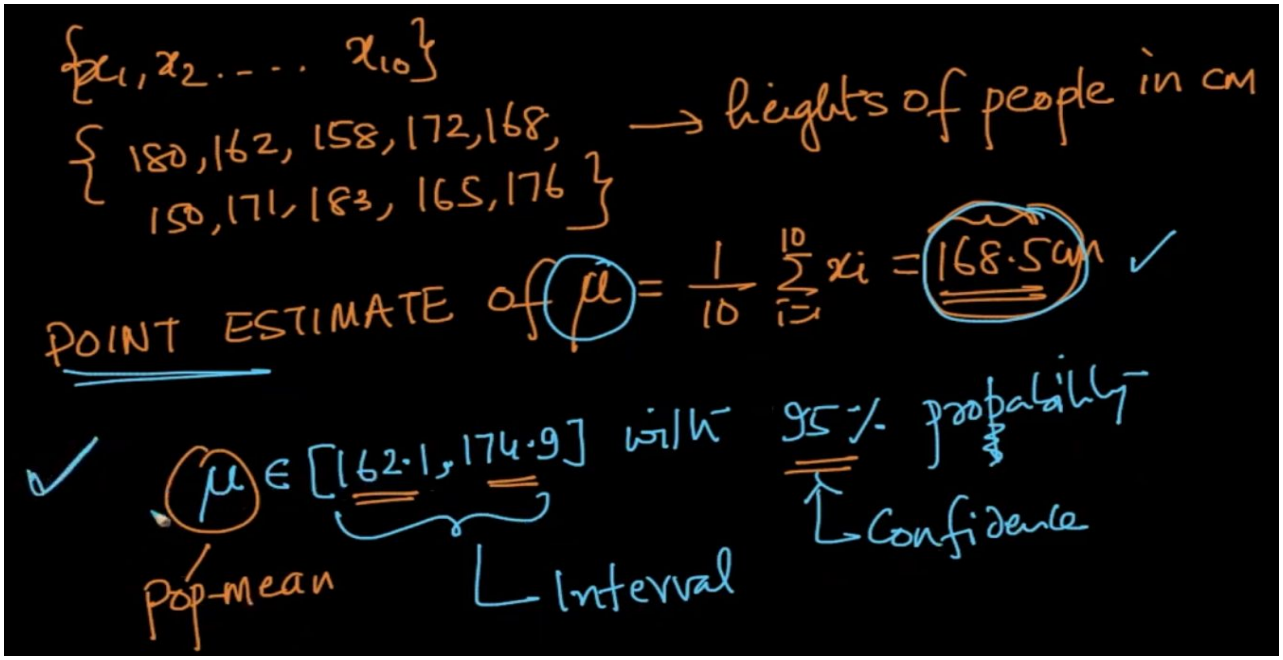
In an e-commerce , if they find that time-spent is +vely correlated money spent as seen then they'll try to design website to increase the time spending(Strategy). Watch 11.23 for more applications

CONFIDENCE INTERVAL



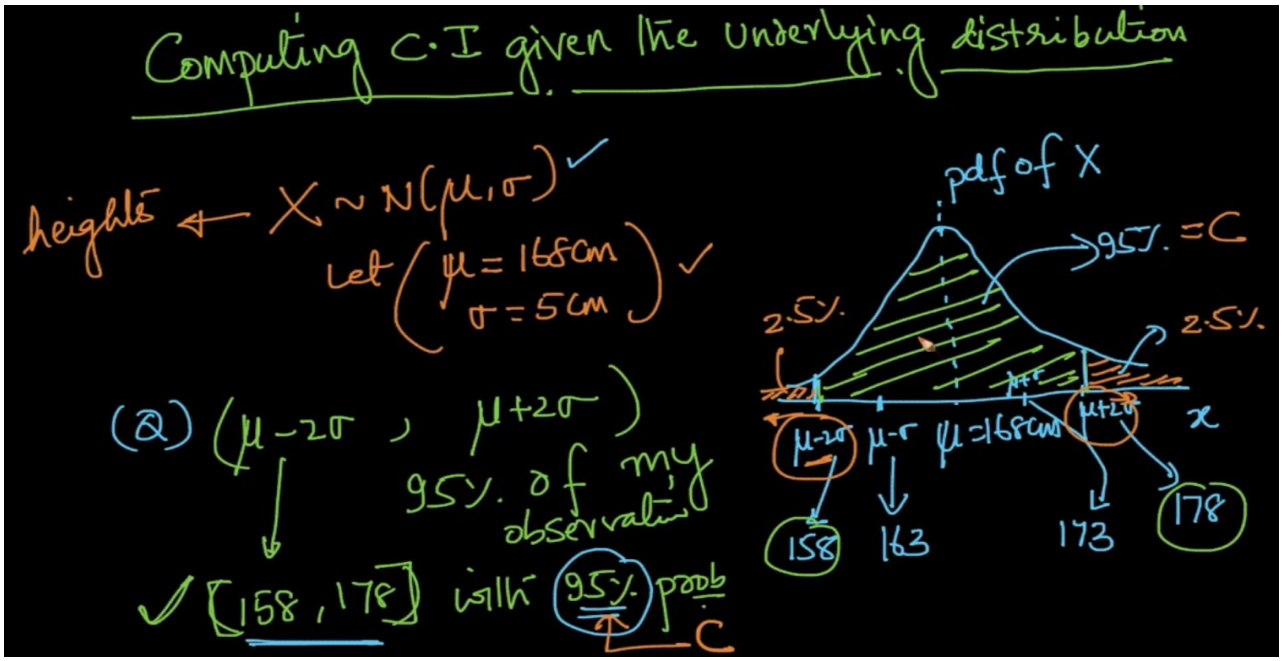
$\bar{x}$ (sample mean) ≈ μ(*pop mean*) which is giving us *point estimate* about how close to our population mean and as n (number of sample) increases we get close to μ (as n↑, ($\bar{x} \rightarrow \mu$))

But there's one better way to estimate this called confidence Interval



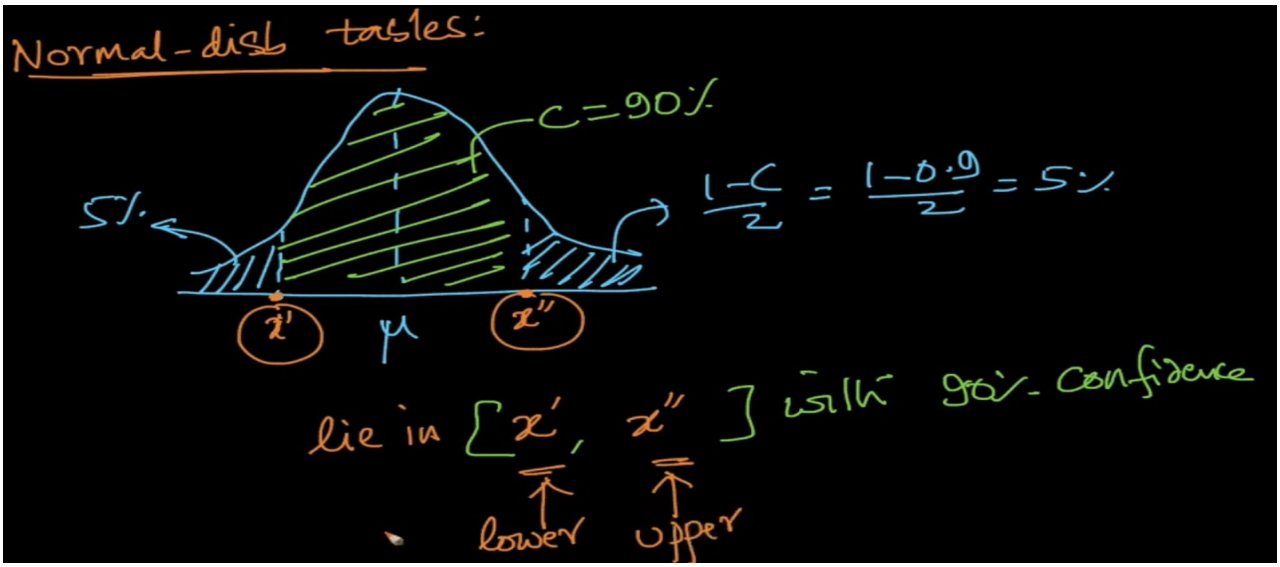
Instead of getting avg we say that there's 95% probability (confidence) that our μ(*pop mean*) lies in this interval above . When we give statements like this we call it Confidence Interval.

CONFIDENCE INTERVAL GIVEN THE UNDERLYING DISTRIBUTION



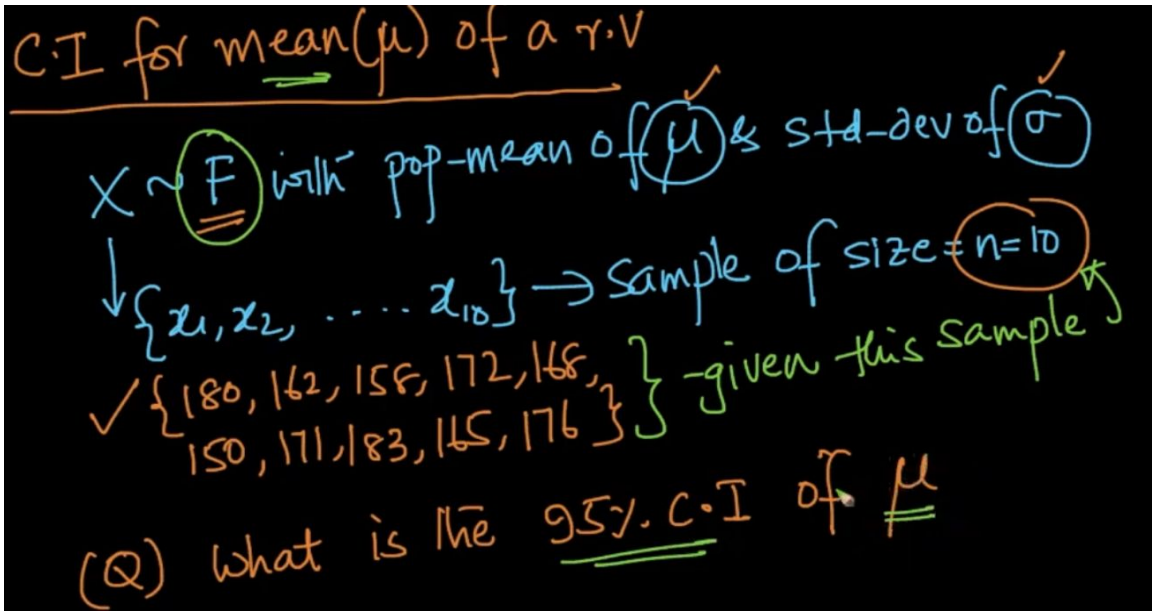
If we are supposed to find the C.I here then we can find it by saying 95% of data(confidence C) lies between ($\mu - 2\sigma, \mu + 2\sigma$) since our distribution is Gaussian and we by using **68-95-99.7 rule**

What if we need to calculate the interval of 90%



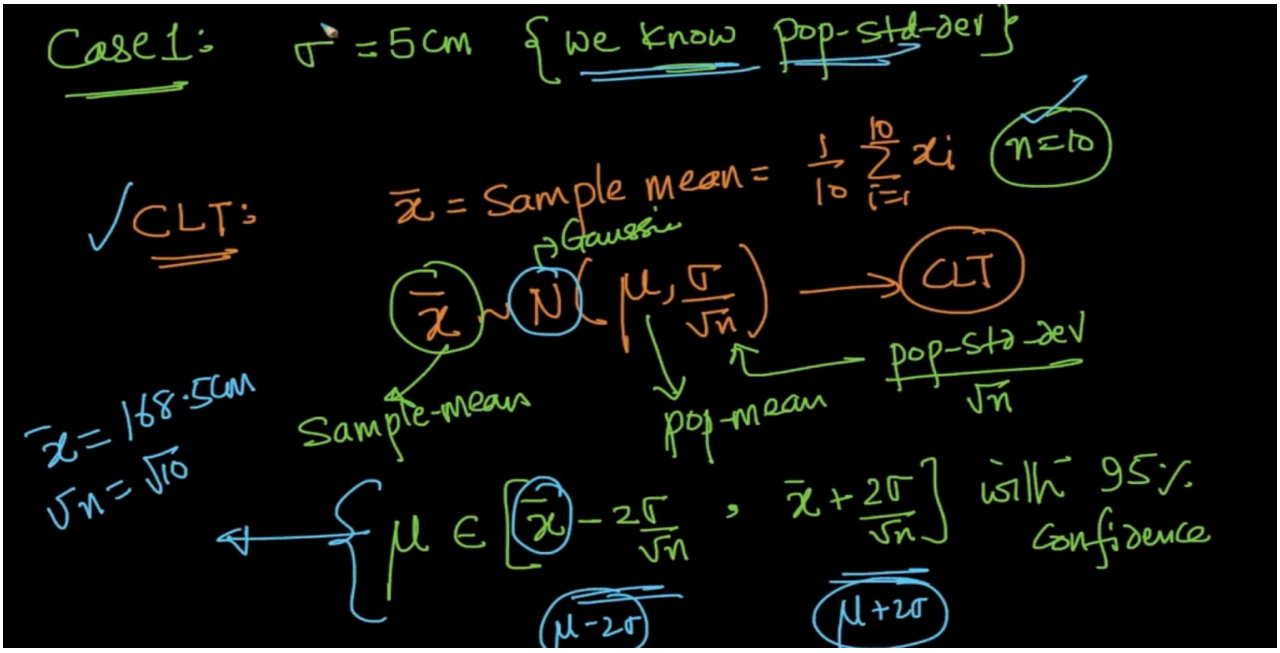
Our C is 90% and the remaining 10% is distributed across ($\frac{1-C}{2}$) then we find the interval by finding the lower bound and upper bound x' and x'' respectively by looking at the N-disb table

C.I for mean(μ) of a random variable



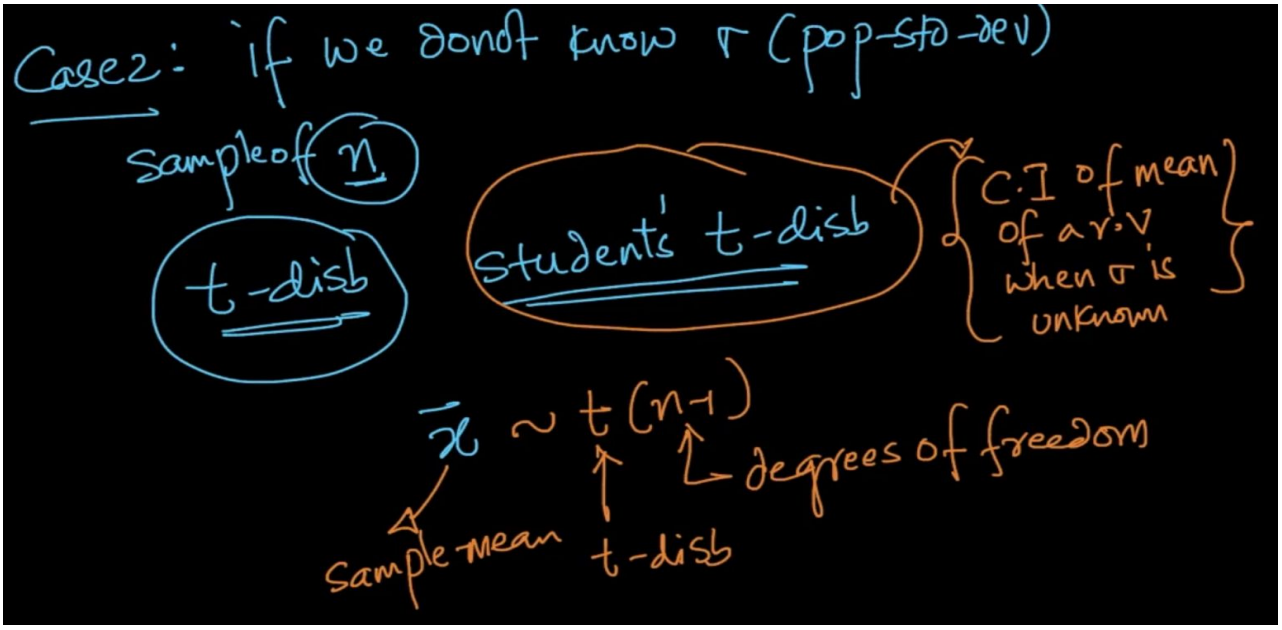
If we are given an r.v with any disb F and finite mean and std.dev then how will we calculate it's C.I with 95% confidence?

Case 1: We know standard deviation

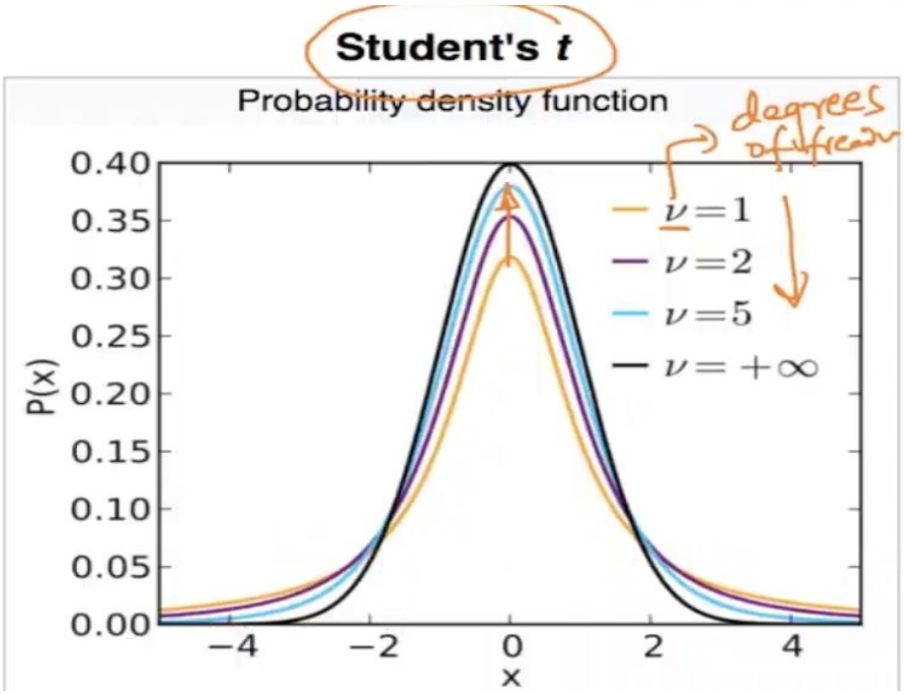


We use Central Limit Theorem(if the μ, σ finite) to convert it into Gaussian and then estimate by using the std devs to estimate 95% of data.

Case 2: We dont know standard deviation



We use student's t-disb for it. It says $\bar{x}$ (sample-mean) ~t(n-1) where t- t-disb and n-1 is degree of freedom .

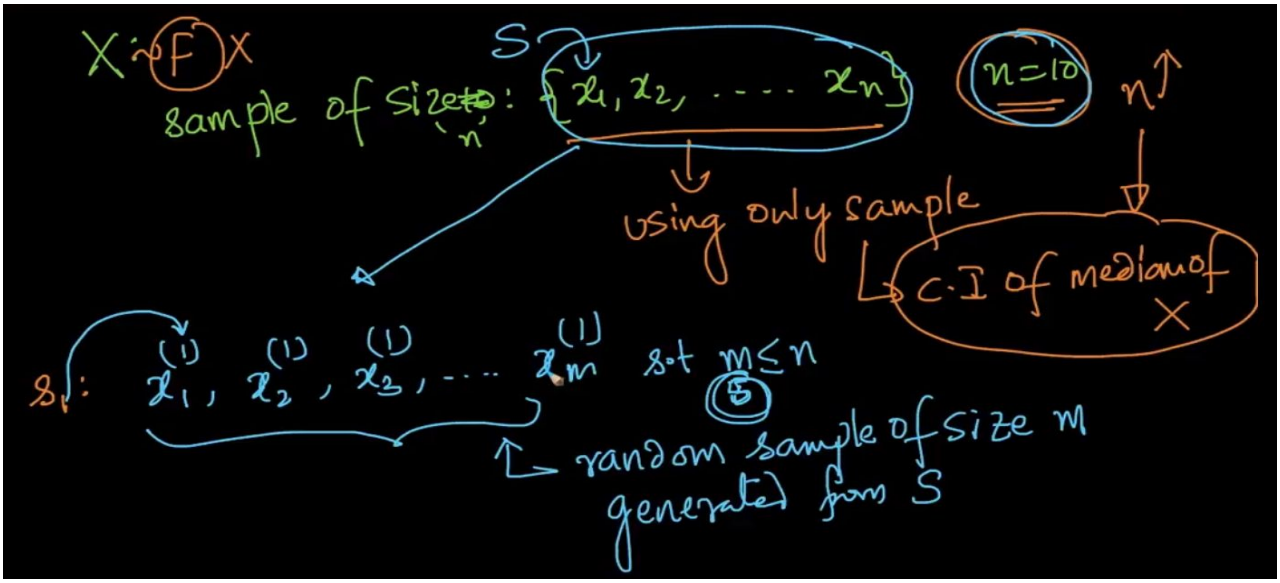


As v (degree of freedom) increases the peakedness increases and with that PDF we can do all the math .

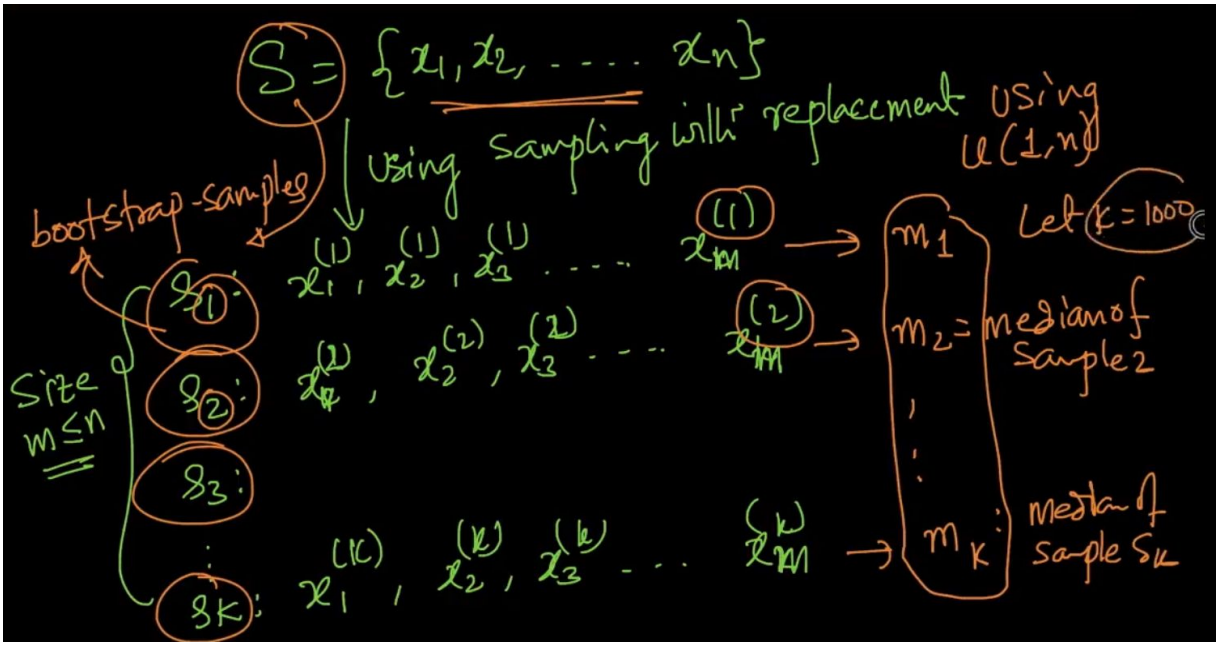
CONFIDENCE INTERVAL USING BOOTSTRAPPING

With the rise of modern computing we can find the C.I of median,percentile, std-dev etc .

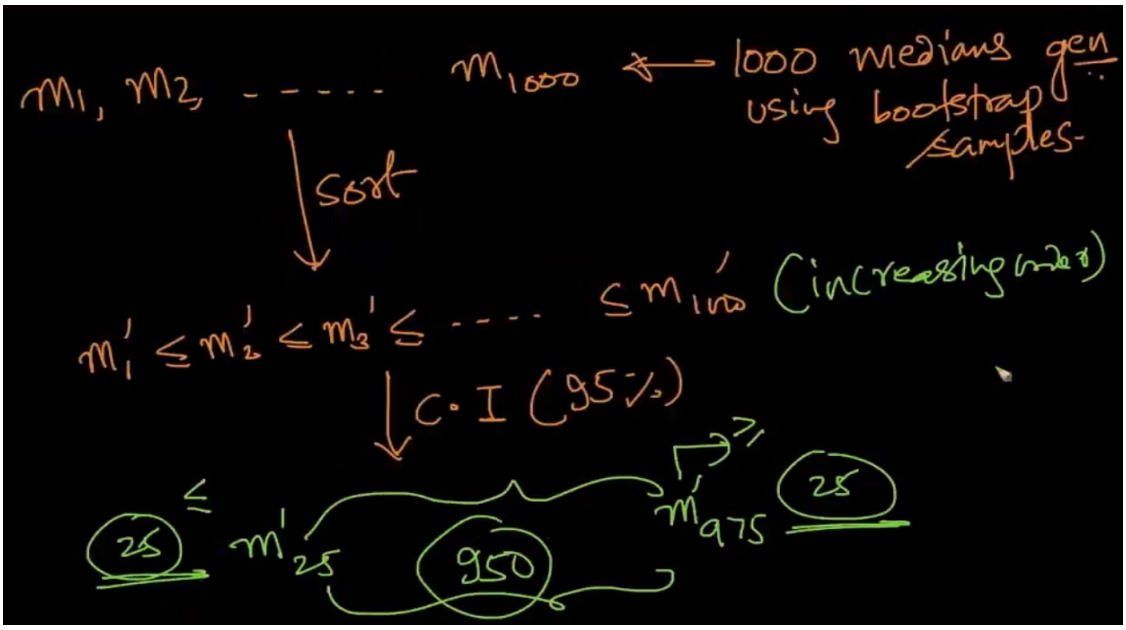
Task- Estimate 95% C.I for median of X.



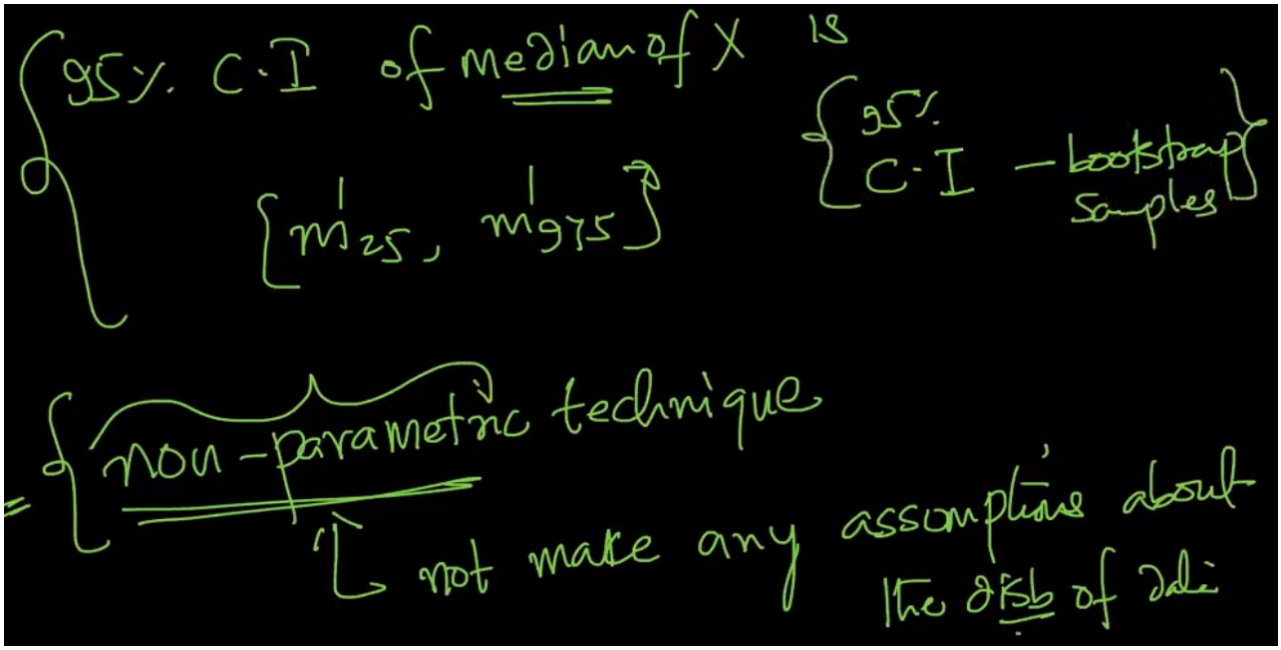
We are getting sample from X and then from sample S we are generating **random** sample of size 'm' ≤ n . Some values can get repeated it is called sampling with repetition.



From sample S we are generating k samples (bootstrap-sample) & then get median of that samples

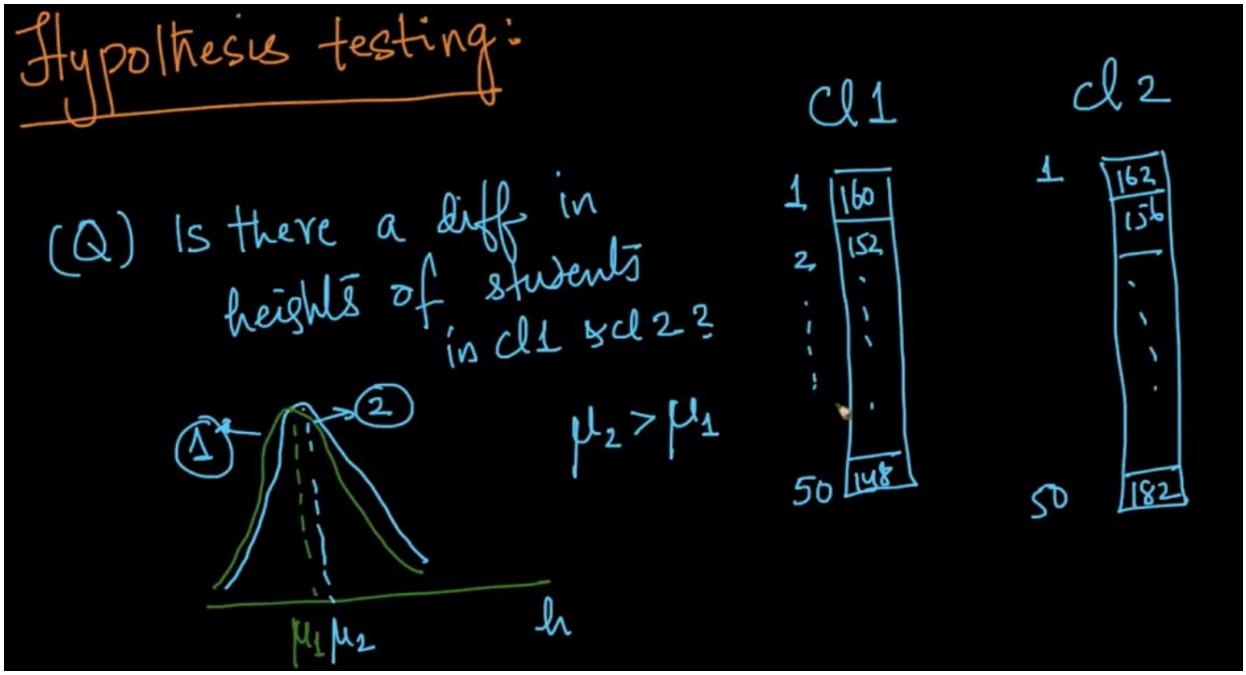


Now after getting medians($m_1, m_2 \ldots . m_{1000}$) of the bootstrap samples we sort them and after that we get the value of m'_{25} *and* m'_{975} since there are 950 values between them which is 95% percent values as k = 1000 where k is number of samples generated.

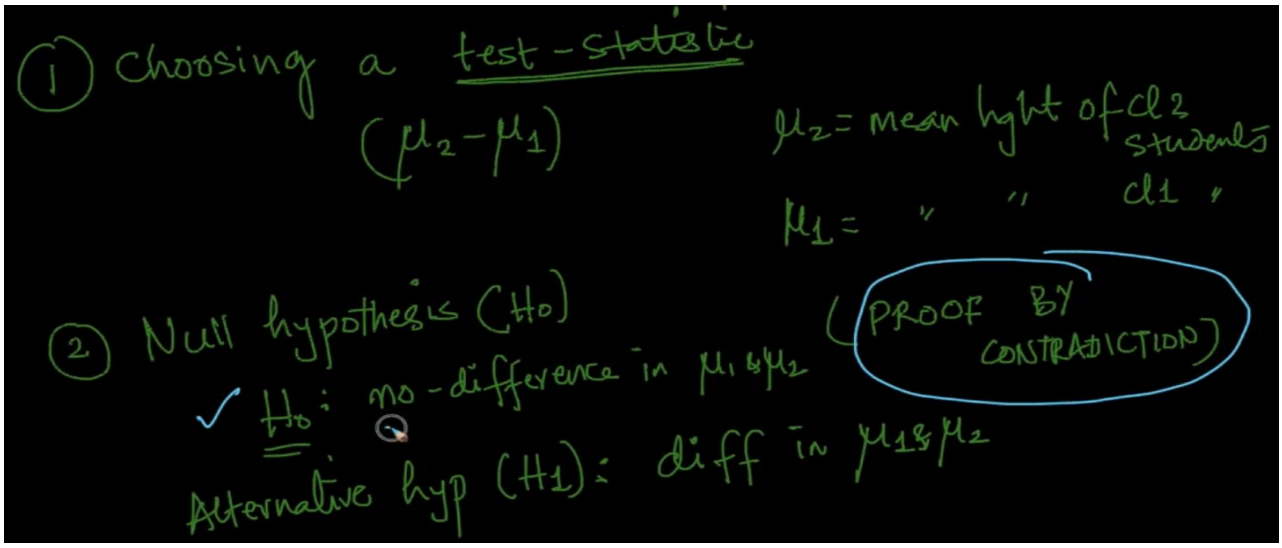


If we want to get C.I of std-devs, percentile , etc then do the steps just like above.This is a non-parametric technique where we dont make any assumptions about the disb

HYPOTHESIS TESTING



We want an absolute answer to the above question. We want to quantify this and there's a method called hypothesis testing to do that. Let's see the steps



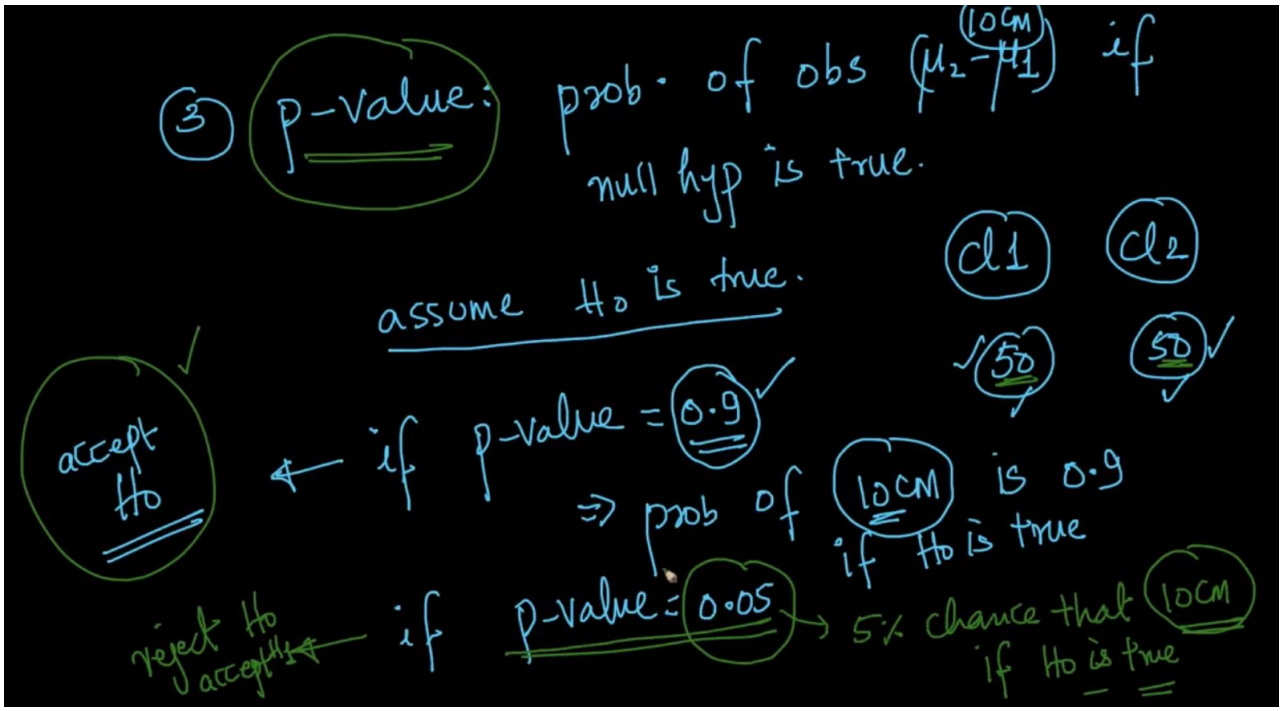
1) We choose a **test statistic** to see if there's a value and the obvious answer here is ($\mu_2 - \mu_1$)

2) **Null hypothesis(** H_o **)** : It says no difference in the means.

Alternative hypothesis (H_1) says that there's difference in means

So we work with proof by contradiction which means we assume our **null-hypothesis** is true but if it's incorrect then it's proved that the hypothesis is incorrect.

3) P-value

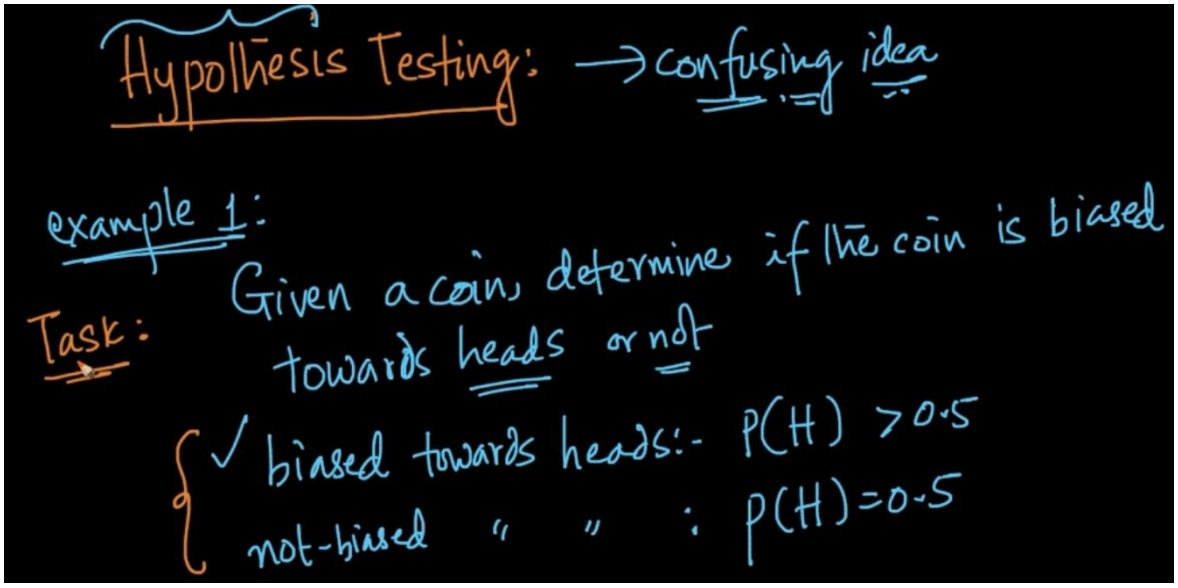


P-value: Probability of observing $(\mu_2 - \mu_1) = 10$ if null hypothesis is true. We assume that H_o is true .

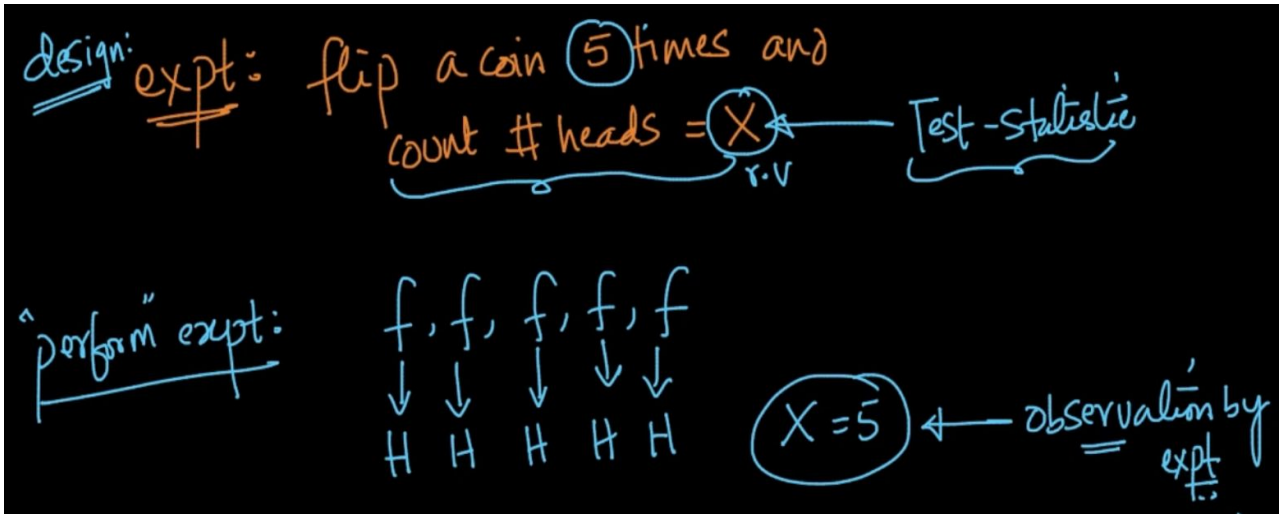
If p-value is high let's say 0.9 then we accept our null hyp(H_o)

If p-value is low let's say 0.05 then we reject H_o or accept H_1

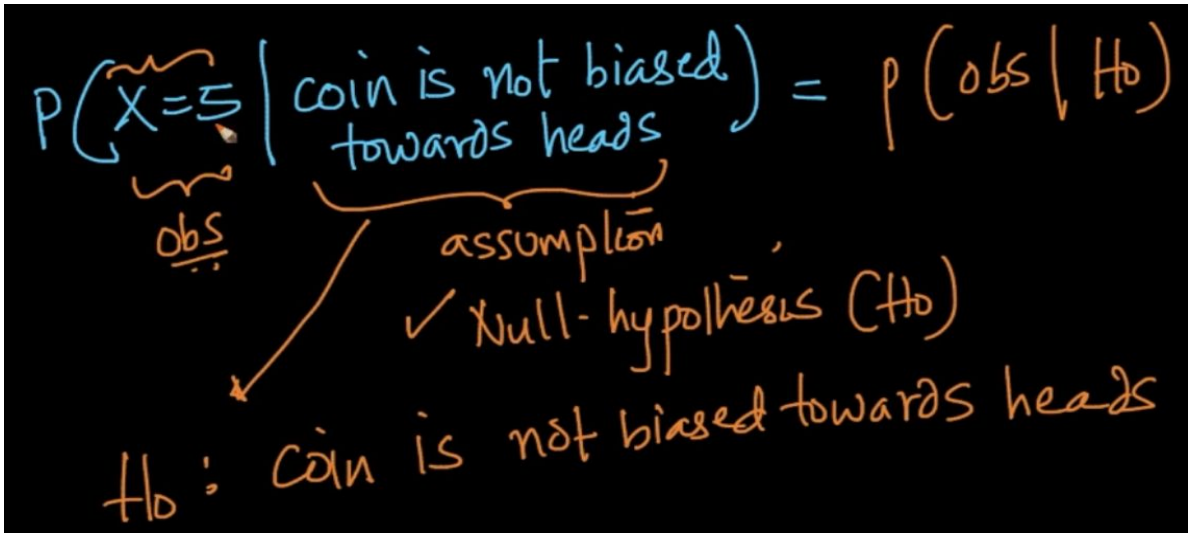
HYPOTHESIS TESTING WITH COIN TOSS INTUITION



If we are saying coin is biased it means that P(That event) > 0.5 since can have only two values heads and tails.

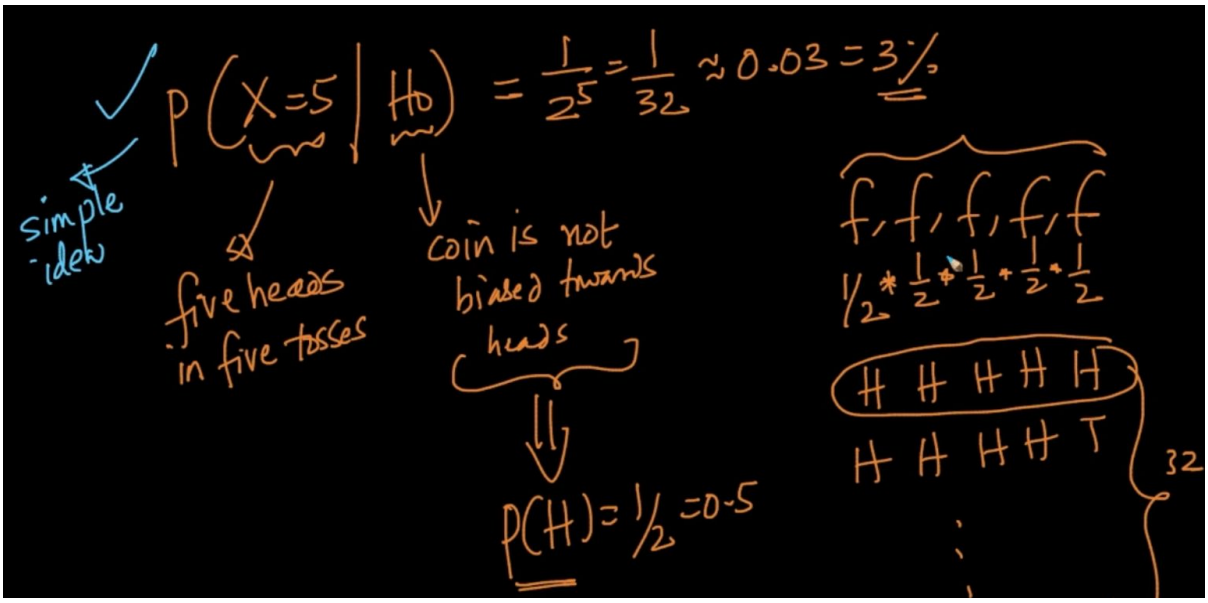


The first step in Hypothesis testing is experimentation. So we design an experiment saying that flip a coins and count no. of heads = X . This X is a random variable called Test-statistic. We flip the coin 5 times and get heads every time. Let's ask a simple question after this experimentation.

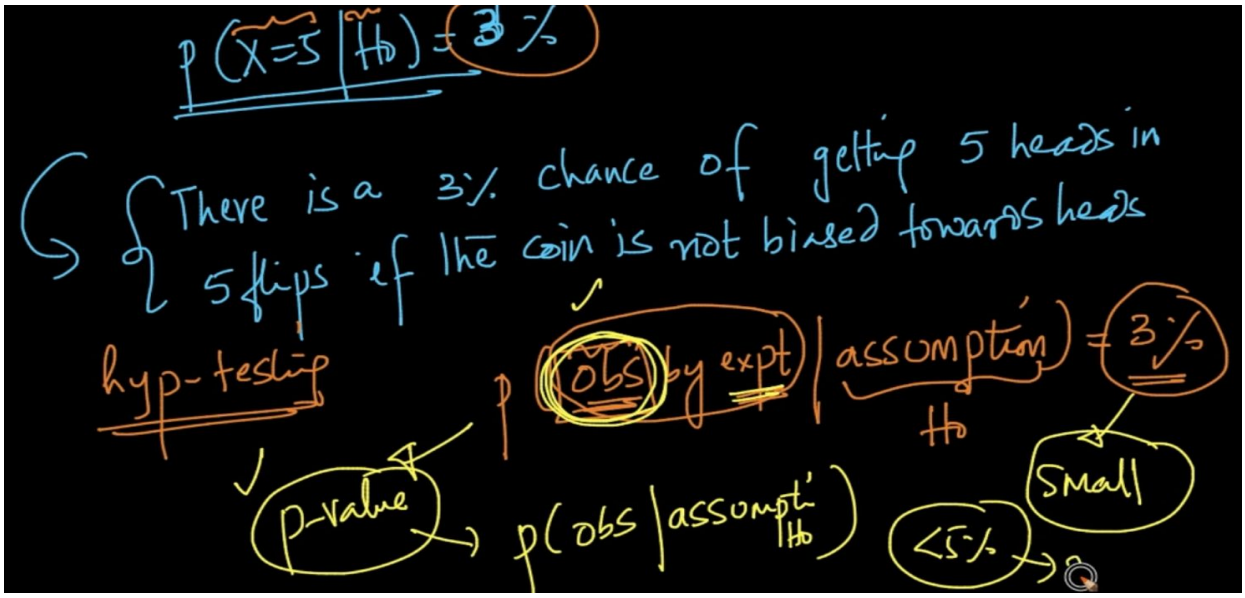


Chance of X = 5 given our assumption (P(X=5 | coin is not biased towards heads) = P(obs/ H_o)

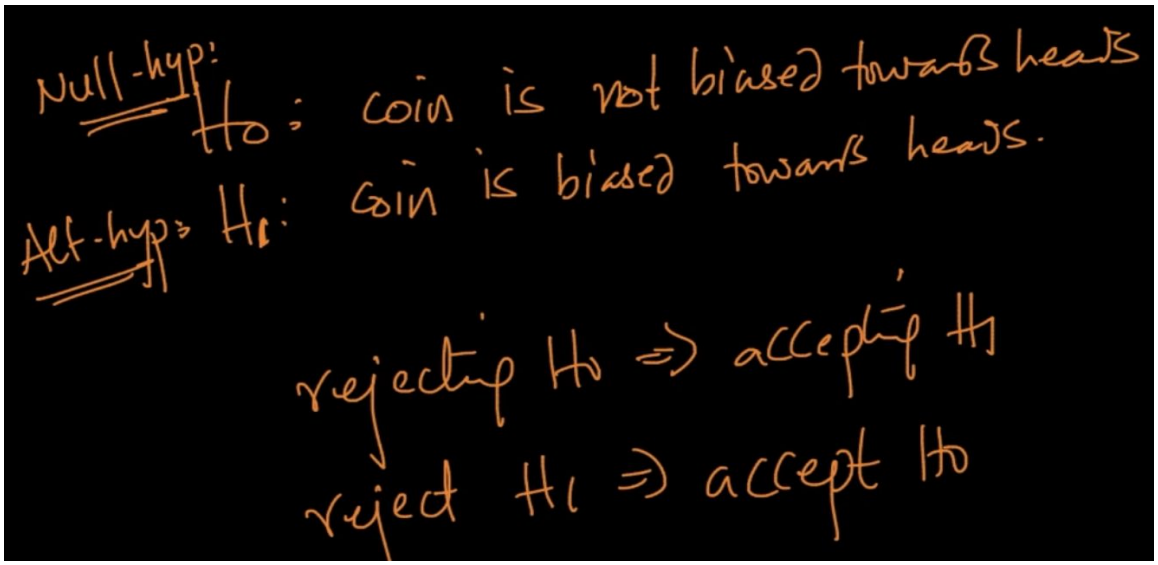
Our assumption is mostly null-hypothesis because it's kinda obvious because normally the coins are not biased.Let's learn how to compute it



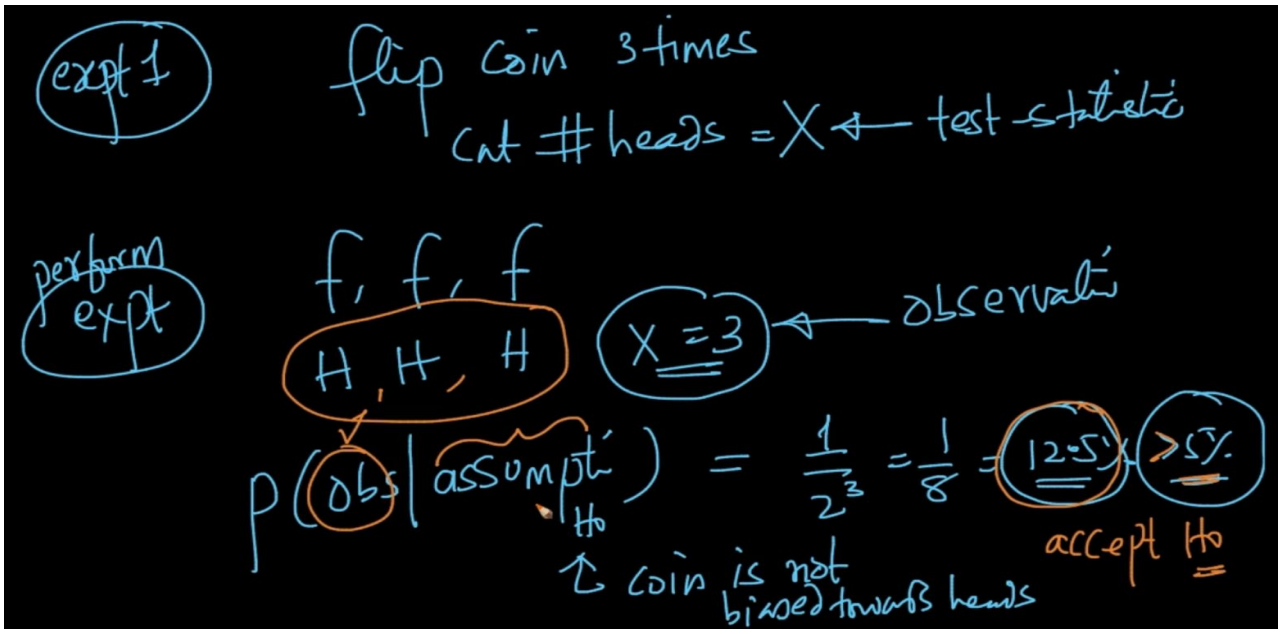
Probability of getting head when a coin is **fair** is ½. We want to know the P(X=5) given coin is fair. So it'll be ½*½*½..½ five times because coin is flipped five times. Which is 3% .Note that we are using very simple ideas for our null-hypothesis (H_o).



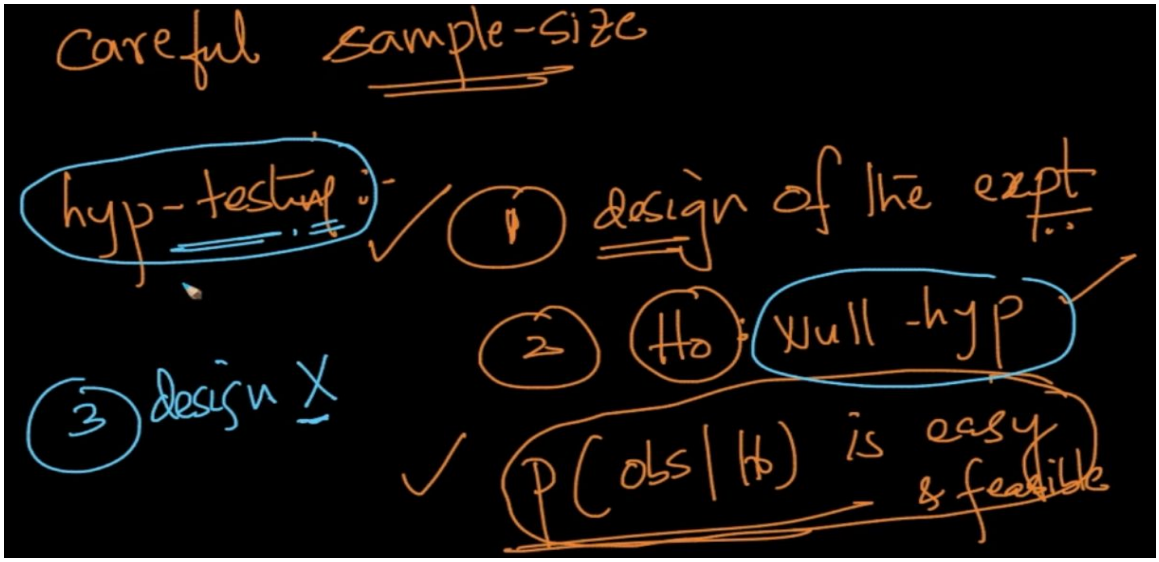
Here, P(Obs| H_0 (assumption)) is called P-value . Here the P-value < 5% then it means that our assumption may be wrong because P is very less and by that we accept our alternate hypothesis/ assumption that Coin is biased or not fair. Idea is mentioned below



We choose a sample of 5 by flipping 5 times . We can change the sample size and our results can vary and change a lot. Let's say we've flipped the coins only 3 times

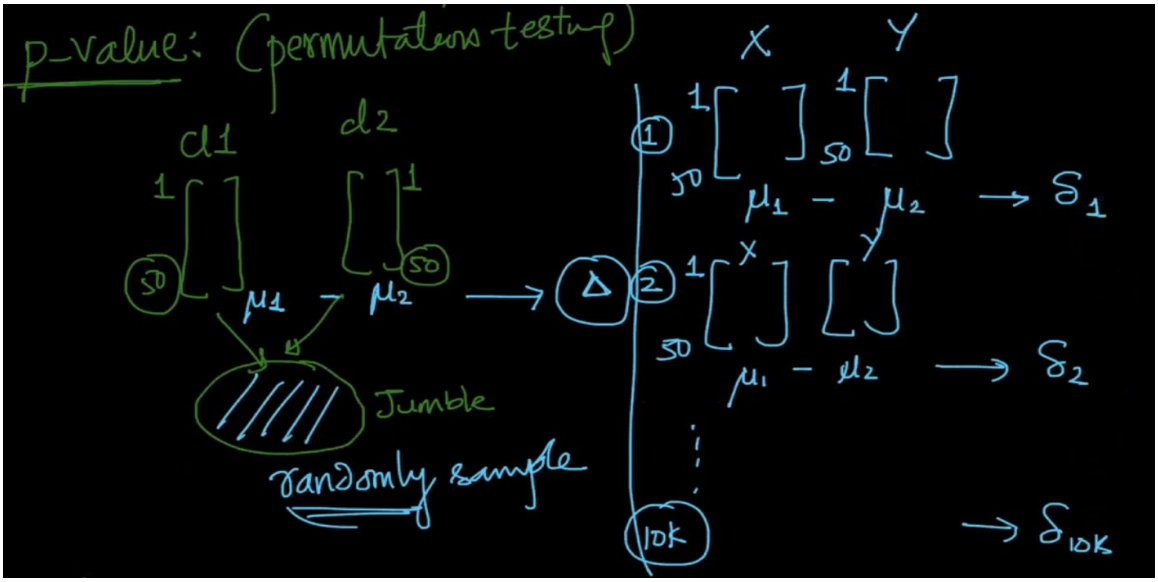


Our P-value is 12.5% and it can't be rejected since 12.5%>5% .SO our assumption that coin is not biased towards heads even after getting 3 heads is valid. So we've to be very careful in choosing the sample-size

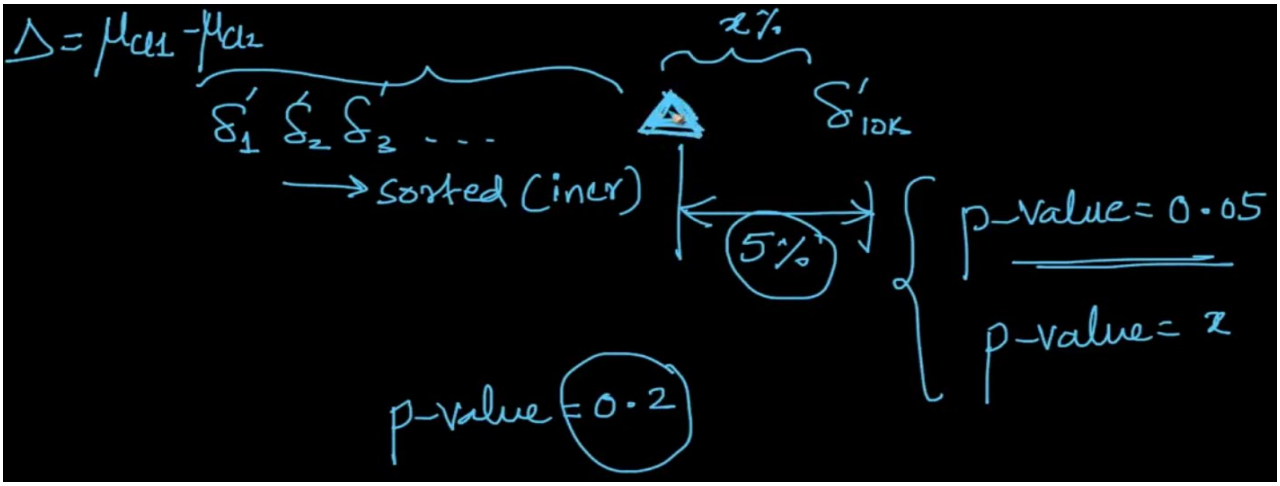


There are three steps : 1) Design of the expt i.e sample size. Expt should be done carefully because if it is wrong then our whole hypothesis testing is wrong. 2) P(obs| H_o) should be easy and feasible 3) Design X (test statistic) properly

CALCULATING P-VALUE



We've two classes Cl1 & Cl2 (heights) of size 50 and calculate $(\mu_1 - \mu_2) = \Delta$ and then take them and randomly jumble them and calculate $(\mu_1 - \mu_2) = \delta$ for 10k times. After that we sort the δ's and see where our Δ is fitting .



If there are 5% of data above let's say $\Delta = 10cm$ then p-value = 0.05 and since it's very small we will reject our H_o (no diff in data) . If p-value = 0.2 then we accept our H_o as p-value > 5%

Kolmogorov - Smirnov test (K-S test)

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 α if

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

Where n and m are the sizes of first and

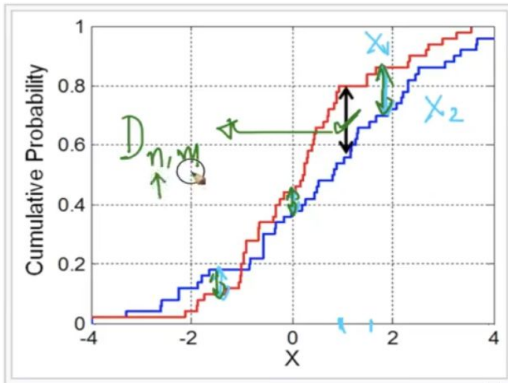


Illustration of the two-sample Kolmogorov–Smirnov statistic. Red and blue lines each correspond to an empirical distribution function, and the black arrow is the two-sample KS statistic.

Let's say our test data X1 & X2 which has size 'n' and 'm' respectively . Now we want to know if they follow the same distribution ? Our null hypothesis is that our X1 and X2 follows same disb

$D_{n,m}$ is our test statistic. It's the max-gap between X1 and X2 as shown in the diag.

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

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

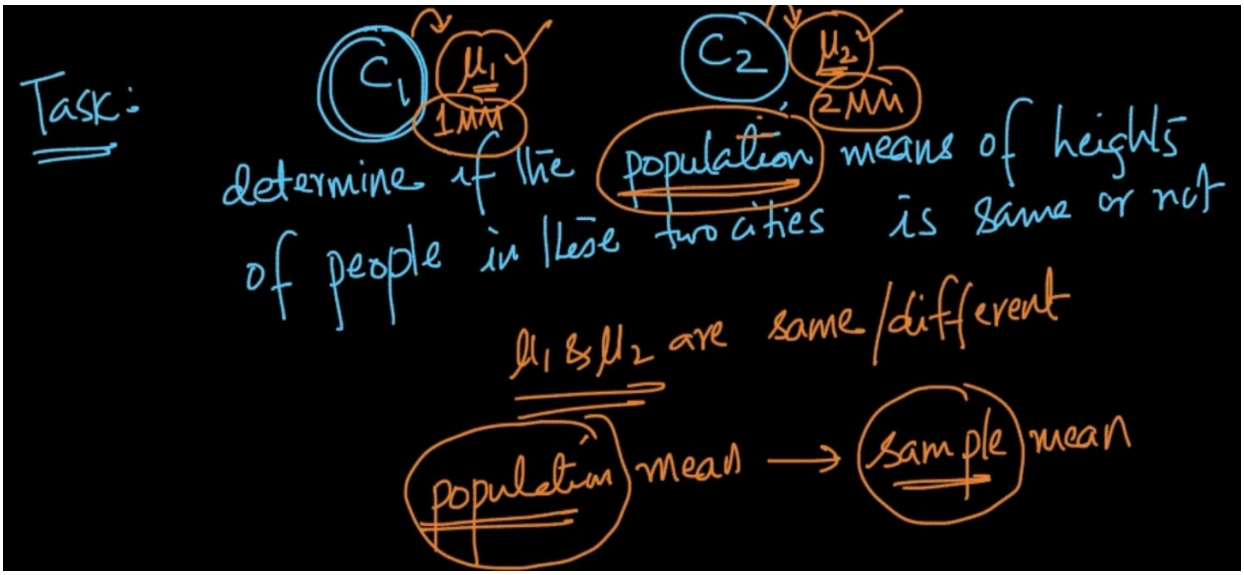
and in general by

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

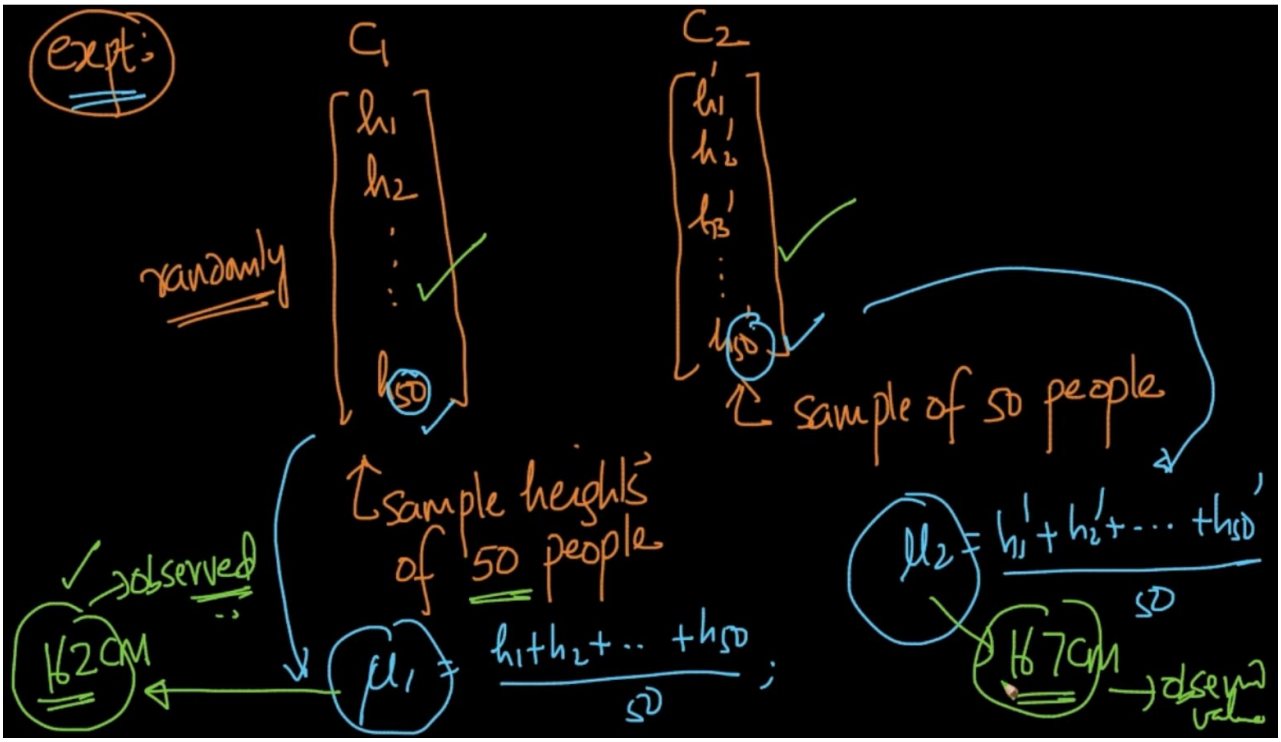
We've a condition for $D_{n,m}$ mentioned above and if it satisfies that condition then our null-hypothesis is rejected. For n = 1000,m=5000, the whole value is 0.047 and $D_{n,m} > 0.047$

(since $D_{n,m} \approx 0.2$)so null-hyp i.e they follow the same disb is rejected at $\alpha/sig.level = 0.05$ .If n=50 and m = 30 then $D_{n,m} < c(\alpha)\sqrt{\frac{n+m}{nm}}$ (0.31) .So we accept our null-hyp at 5% sig-value/p-value

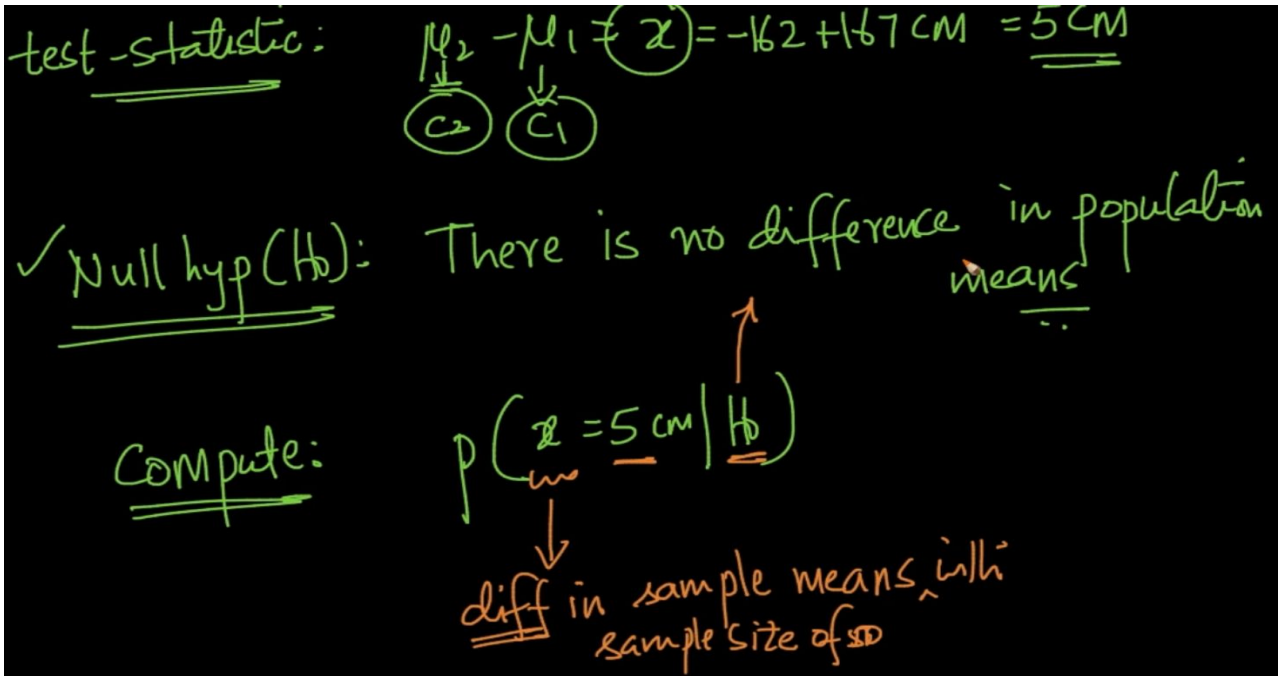
HYPOTHESIS TESTING WITH ANOTHER EXAMPLE



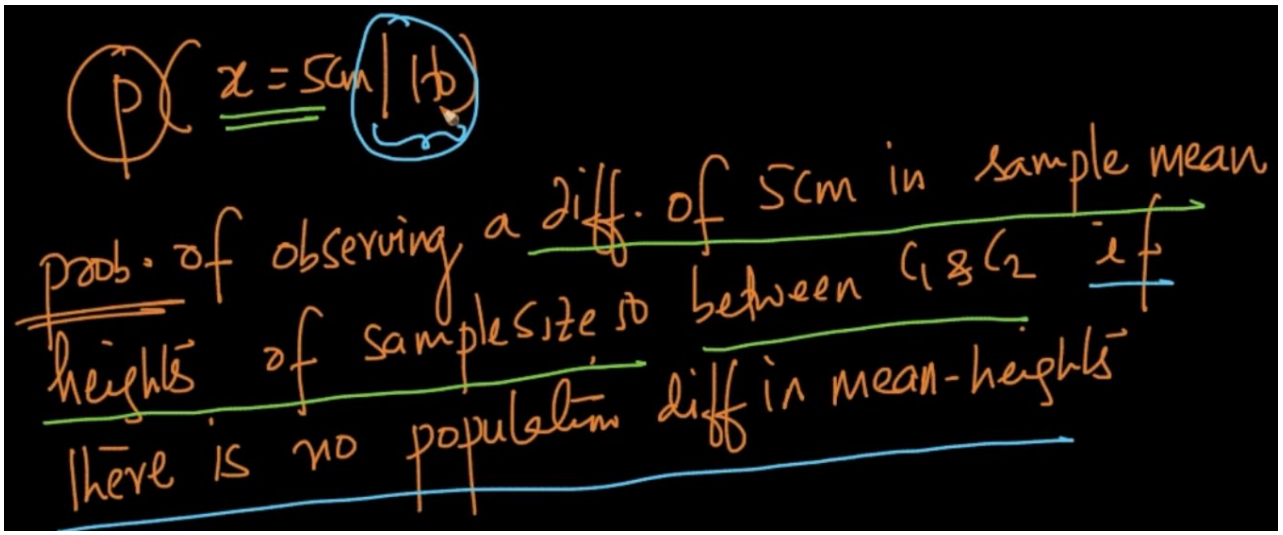
We are taking the sample mean since population mean is obviously expensive.



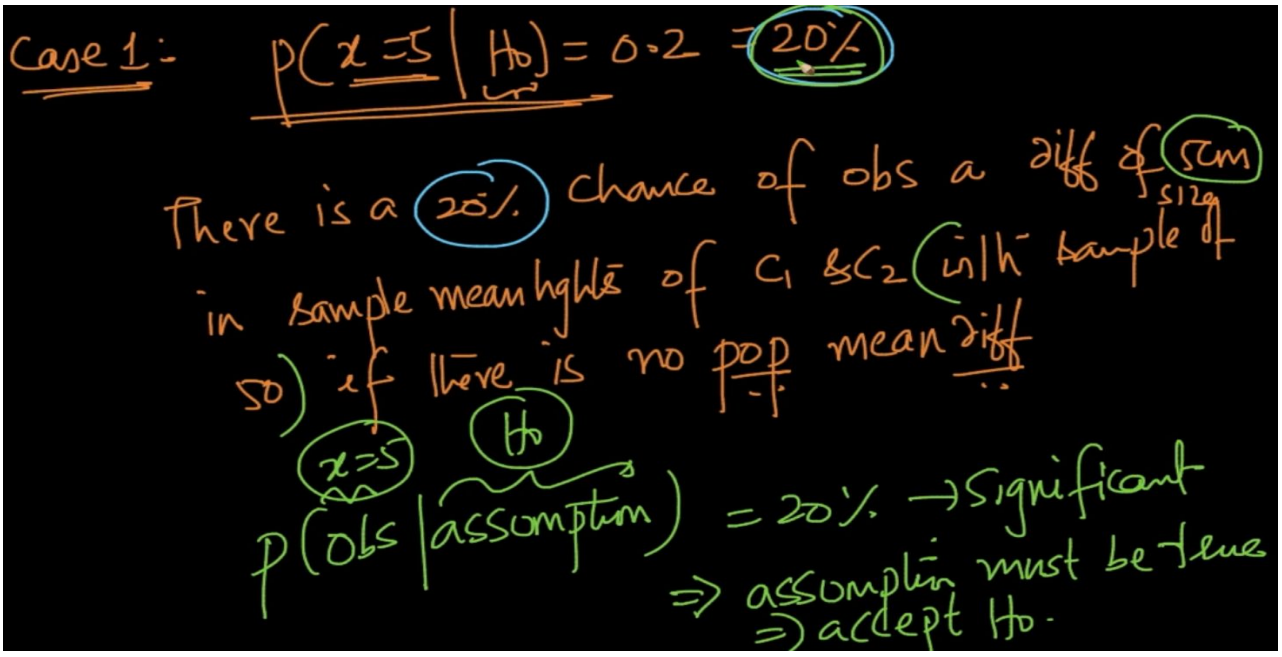
After getting samples we calculate their respective means



The compute part is explained in better English in the next img but basically it is the Probability of observing a difference(value which is 5) in means if null hyp is true.

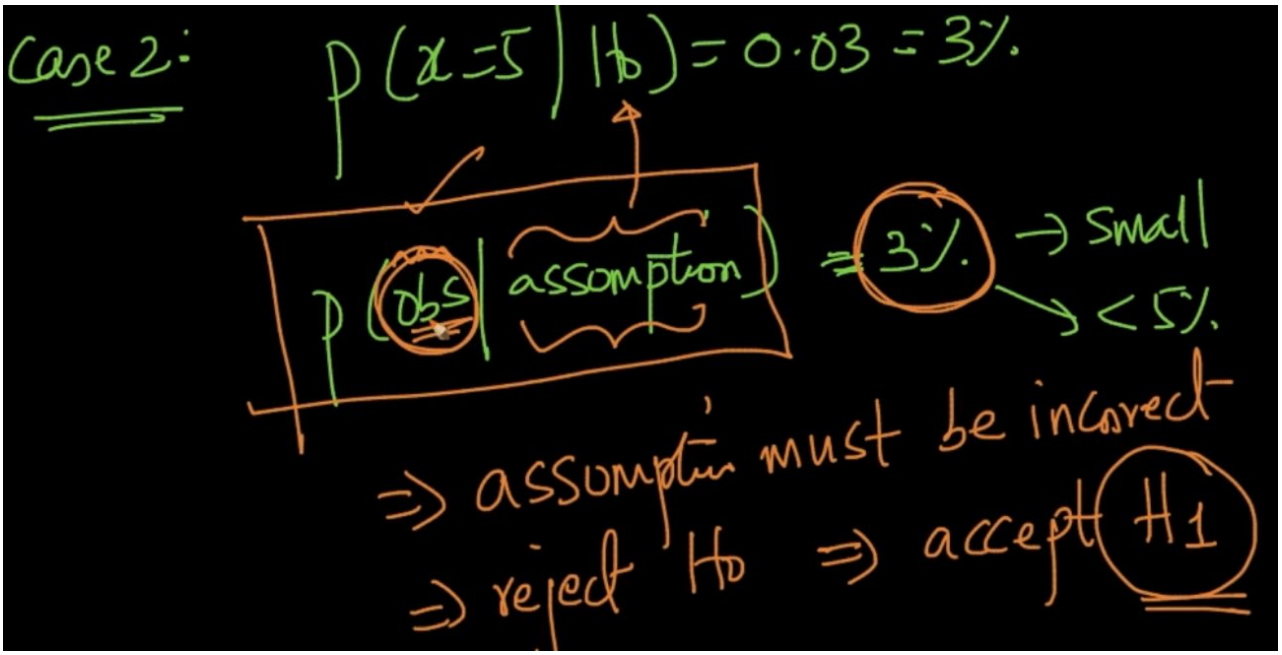


CASE 1 : P-value is 20% or 0.2



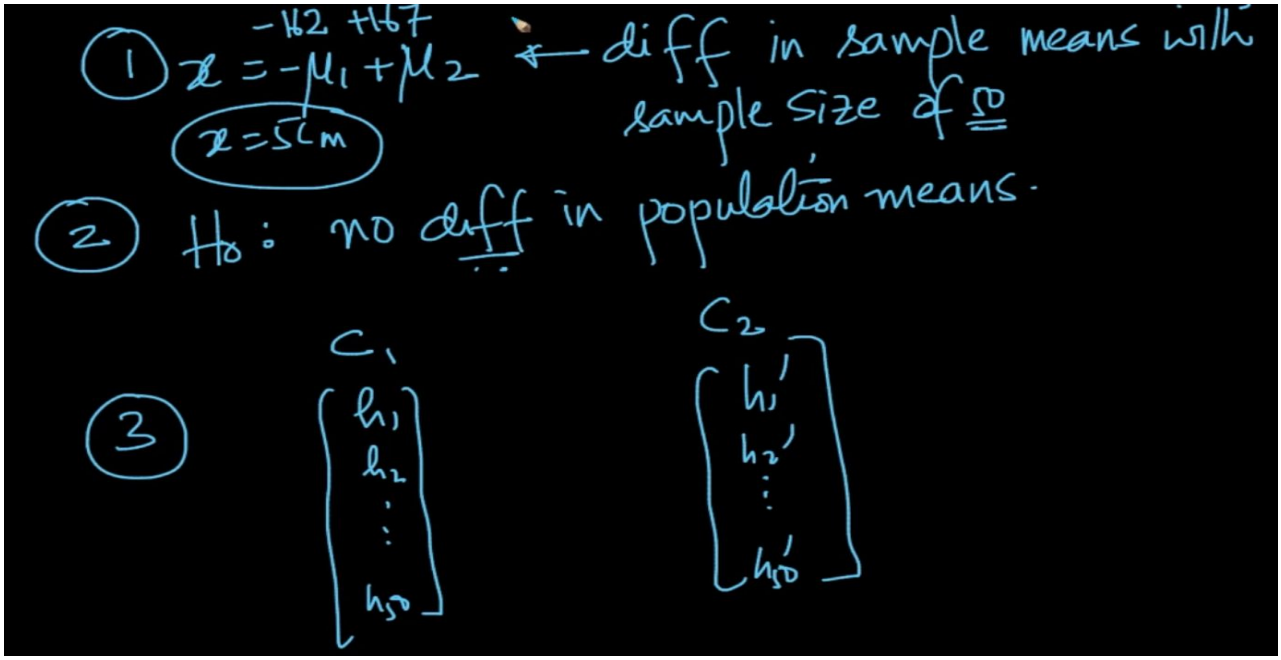
Since 20% is a significant number we'll accept our null hypothesis

CASE 2: P-value is 3% or 0.03

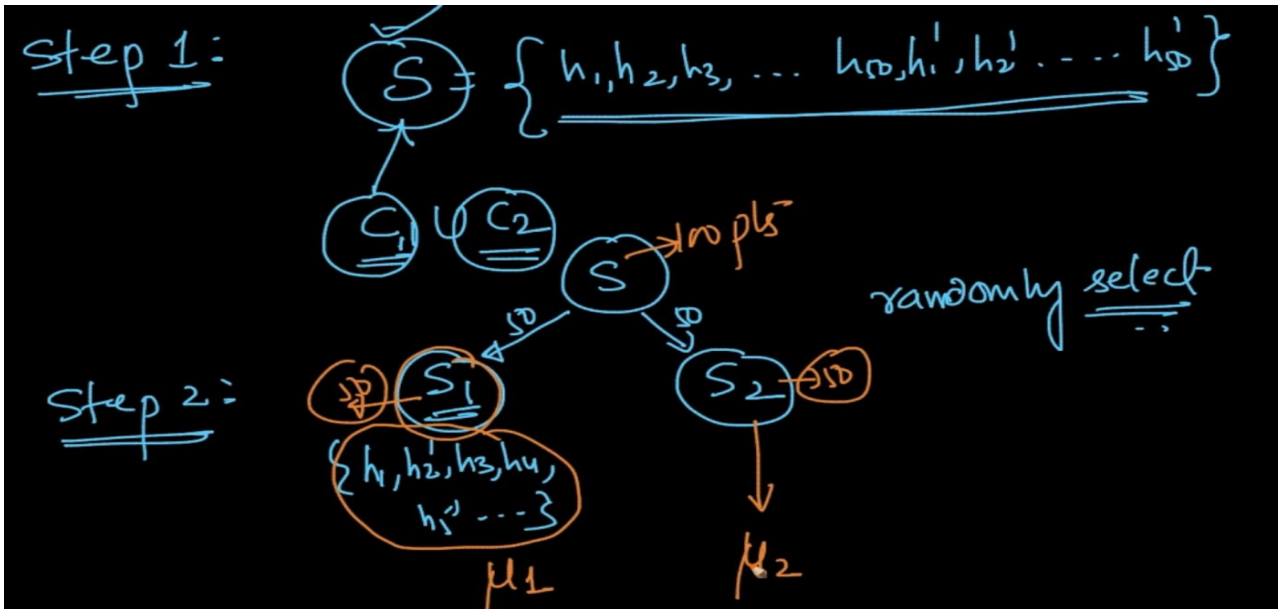


Since P-value is small we reject our assumption and take the alternative assumption/hypothesis(H_1). Note: Our observations are never wrong so only the assumptions can be wrong.

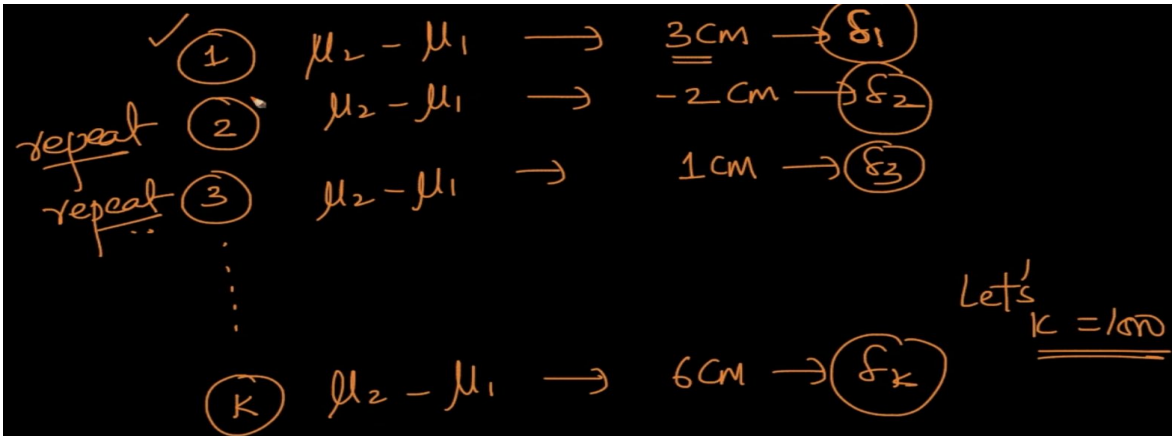
COMPUTING PROBABILITY



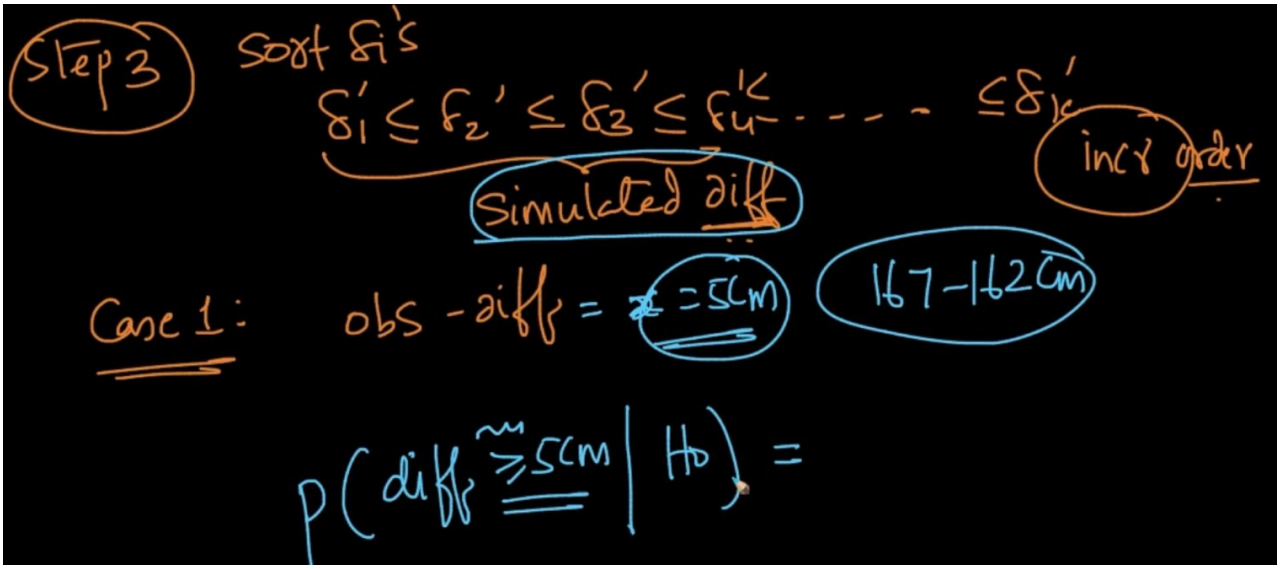
With these 3 points we'll need to compute P-value.



We create a set of 100 points with both samples of C1 and C2. Then we create another sample of size 50 and 50 and get their means μ_1 & μ_2 .This is called resampling (2nd step).So we are calculating the diff in means .So if our null-hyp is correct then there should be no significant difference in means.We are trying to simulate H_0

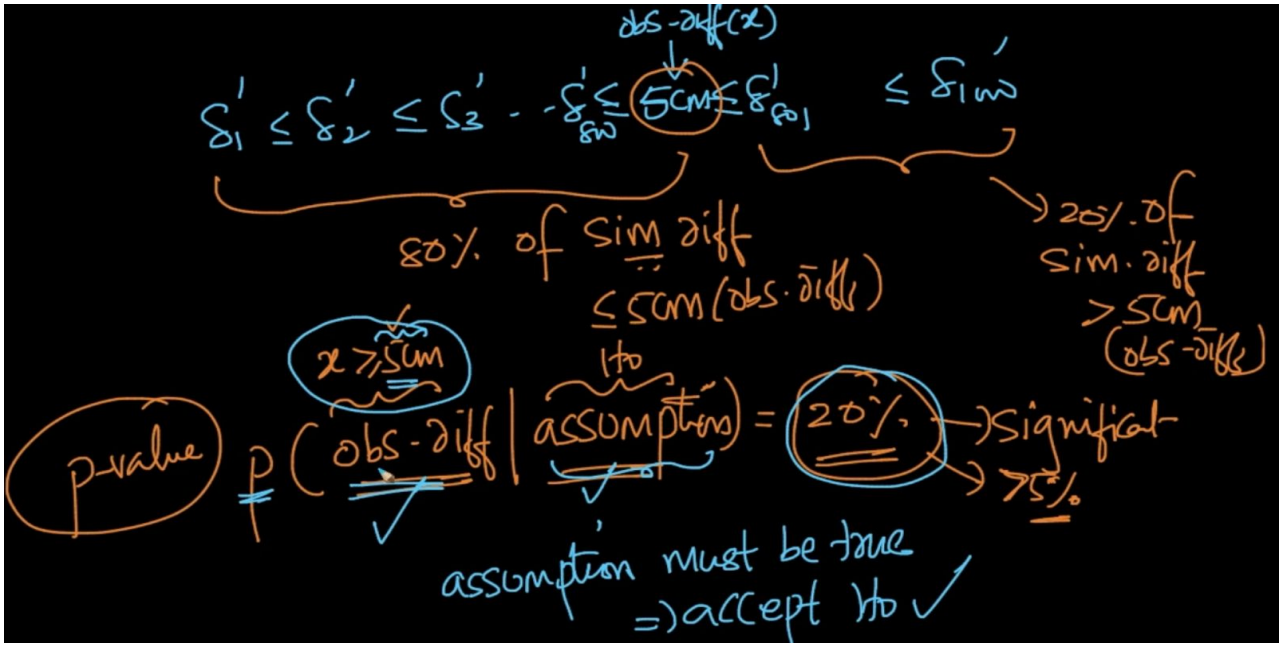


We do all the above two steps for k times (let k = 1000)



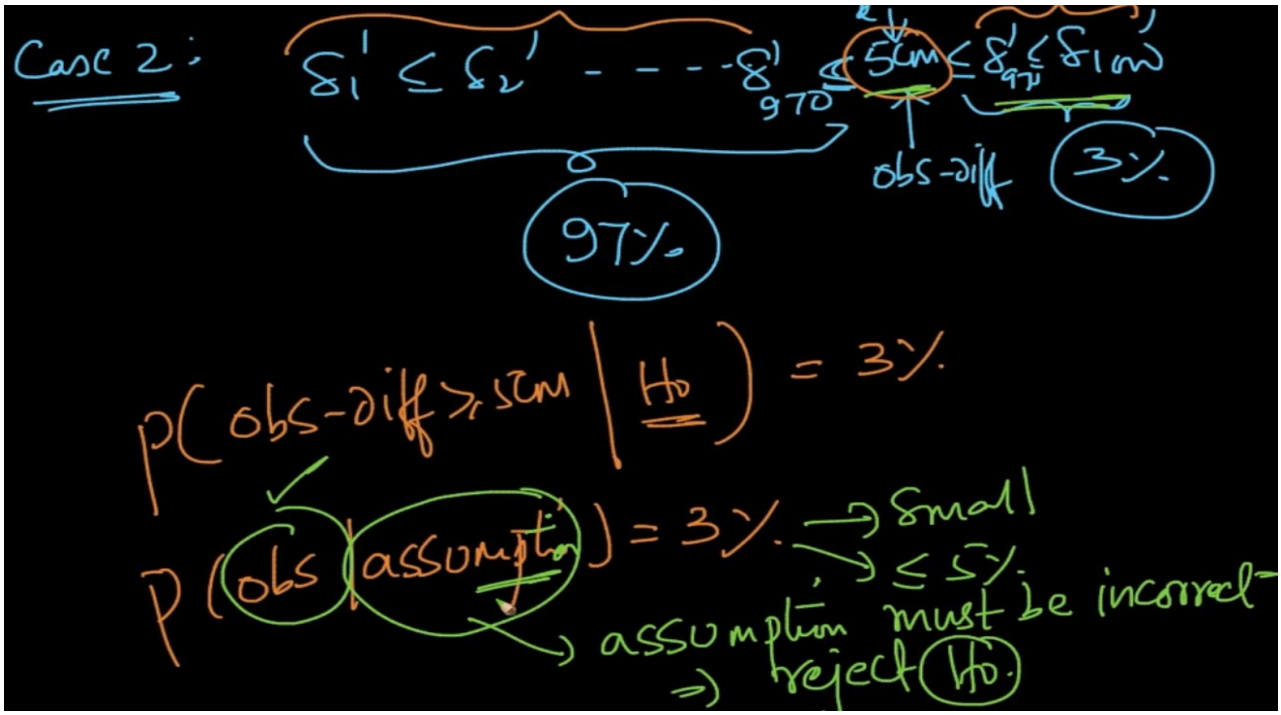
We sort the simulated differences .

Case 1 : P(diff $\geq 5cm \mid H_o$) = 20 %



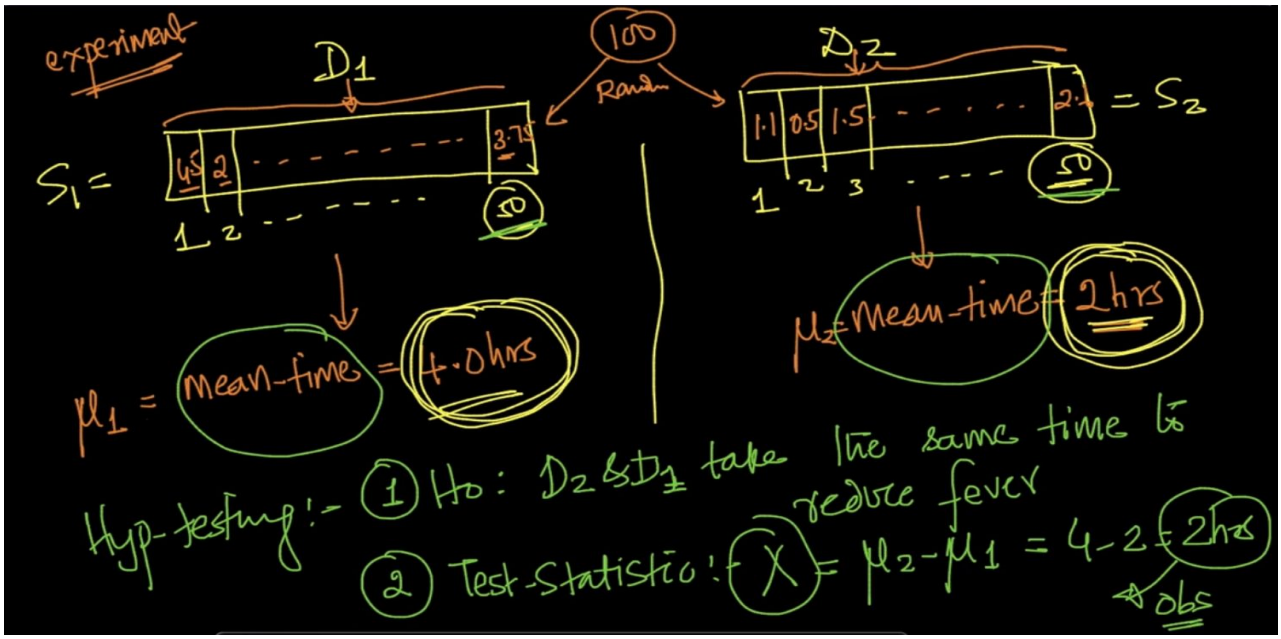
Our observed diff is 5 cm placed at the 800th position. Therefore 20% of values > 5 c.m so that's our P-value and we accept our hypothesis.

Case 2 : P(diff $\geq 5cm \mid H_o$) = 3 %

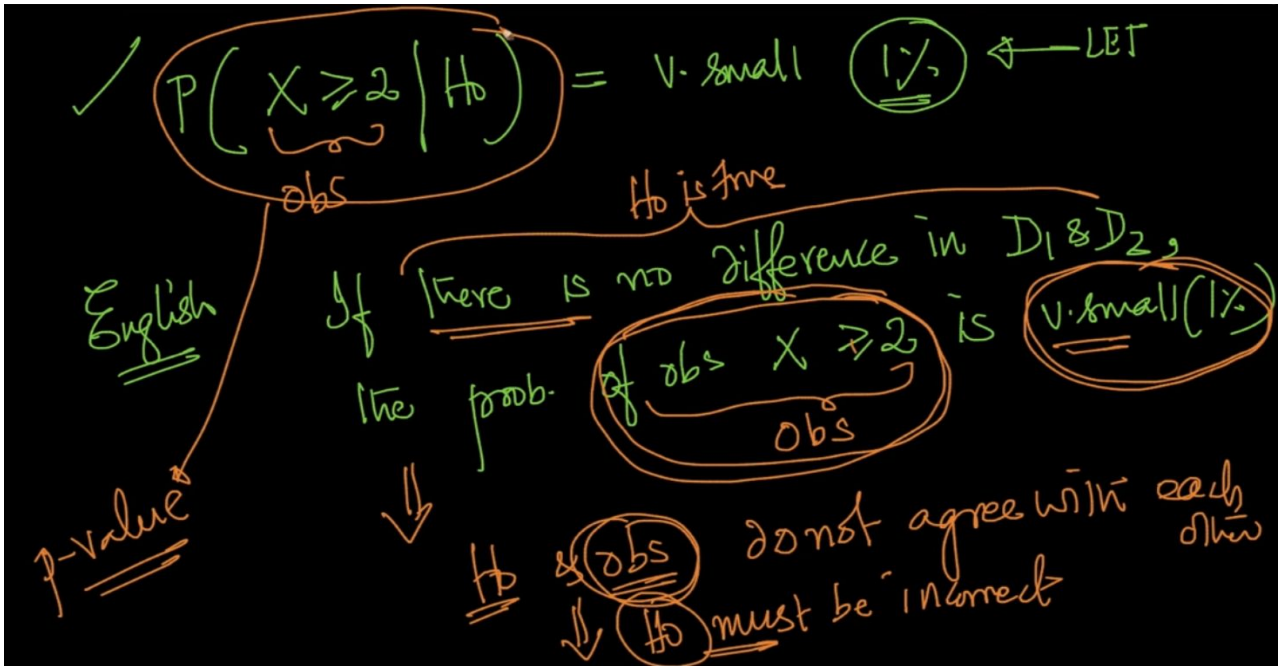


HOW TO USE HYPOTHESIS TESTING

Our task is to determine whether drug D2 takes less time than D1.



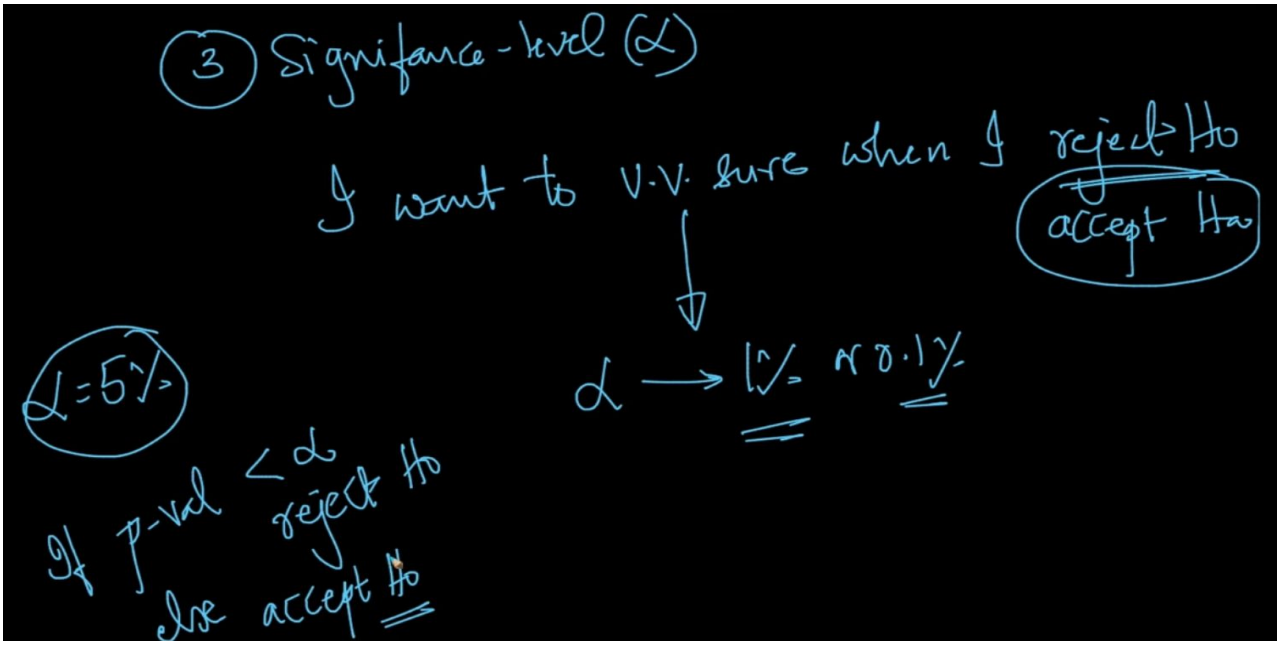
We'll take 100 people and then divide the sample into 50 each . Then we test the drugs on each.



If P - value is very small . If there's no diff in D1 and D2(H_0) the prob of X ≥ 2 *is v.small*(*obs*) then H_o & obs don't agree with each other there we dont accept our H_o

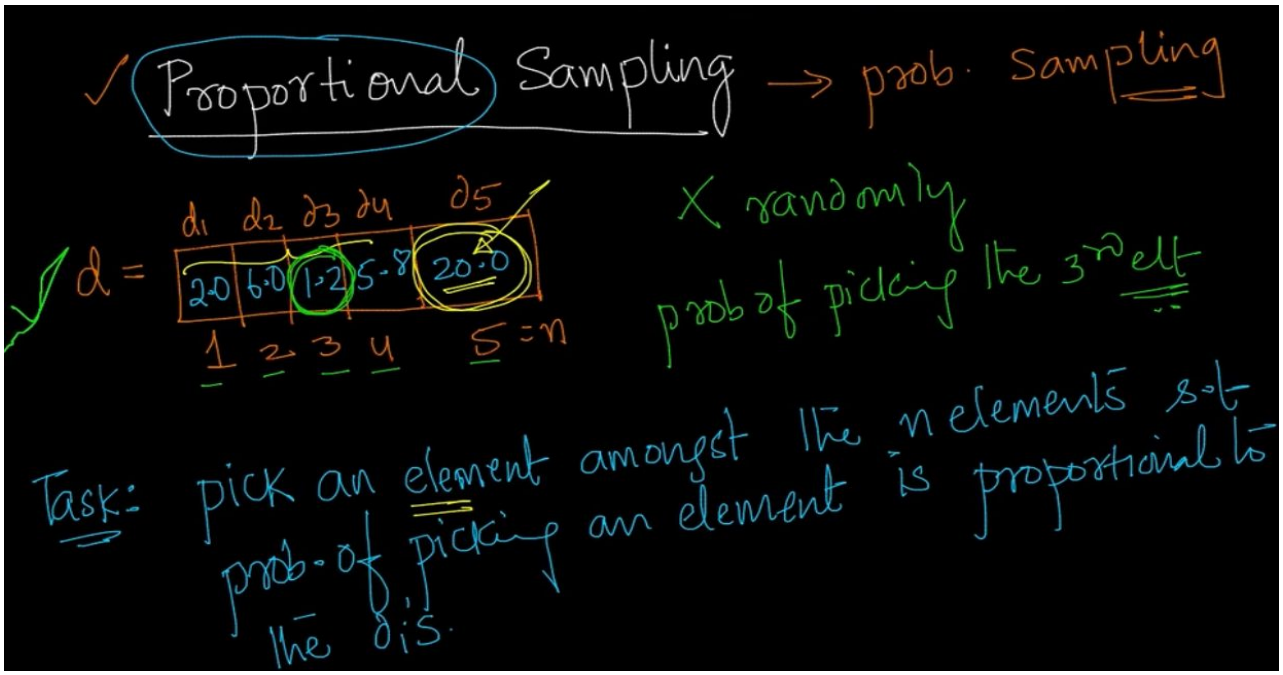
How do we calculate the P-value ?

- By resampling method used above i.e calculating means of different sample ‘k’ times and then observe the differenec

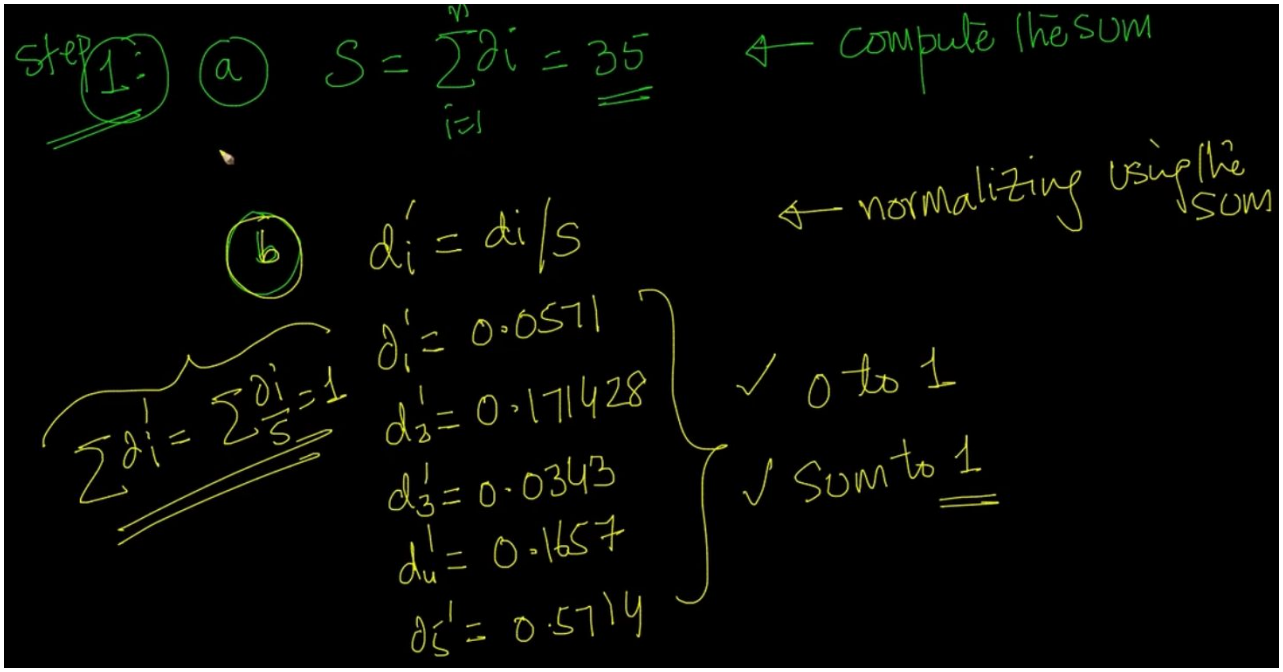


We set a significance level(α) and if our p-value is less than that then we reject our hypothesis

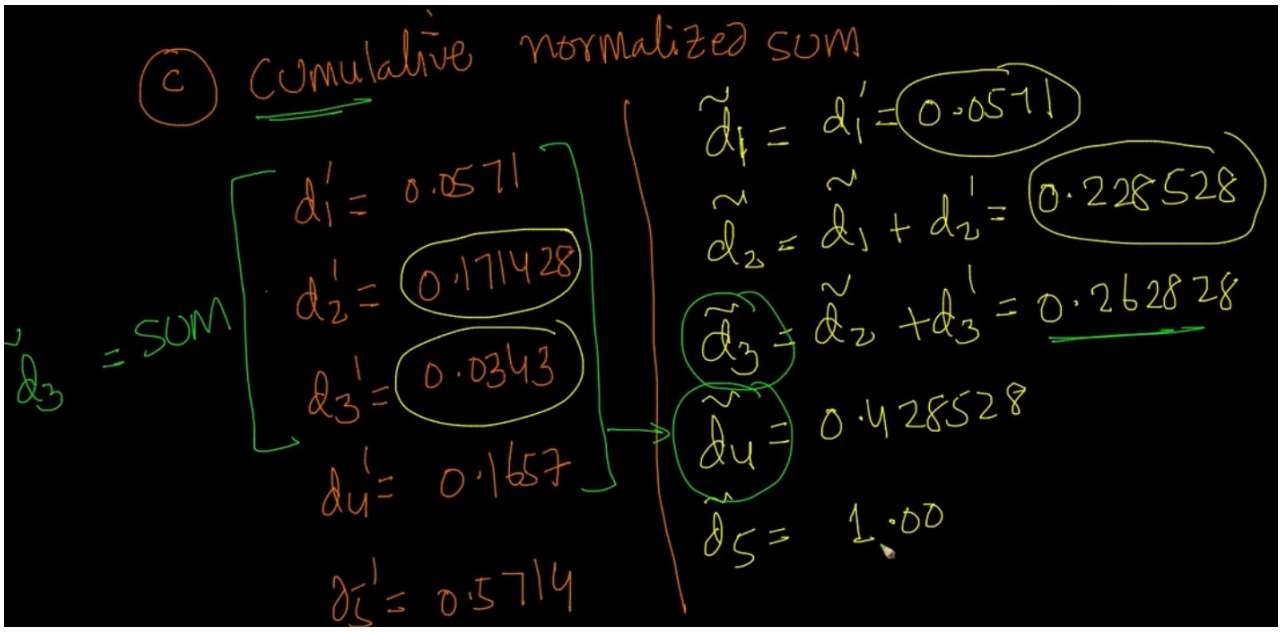
PROBABILITY SAMPLING



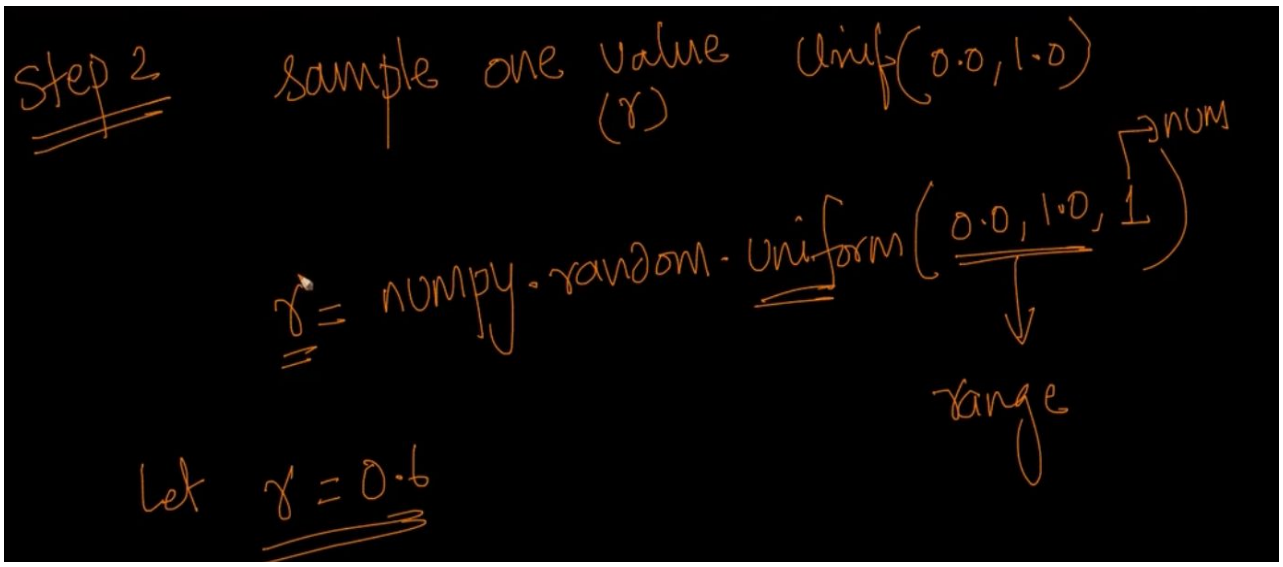
But we don't want it to be random . We want the 20 to be more likely picked up and 1.2 should be the least likely.

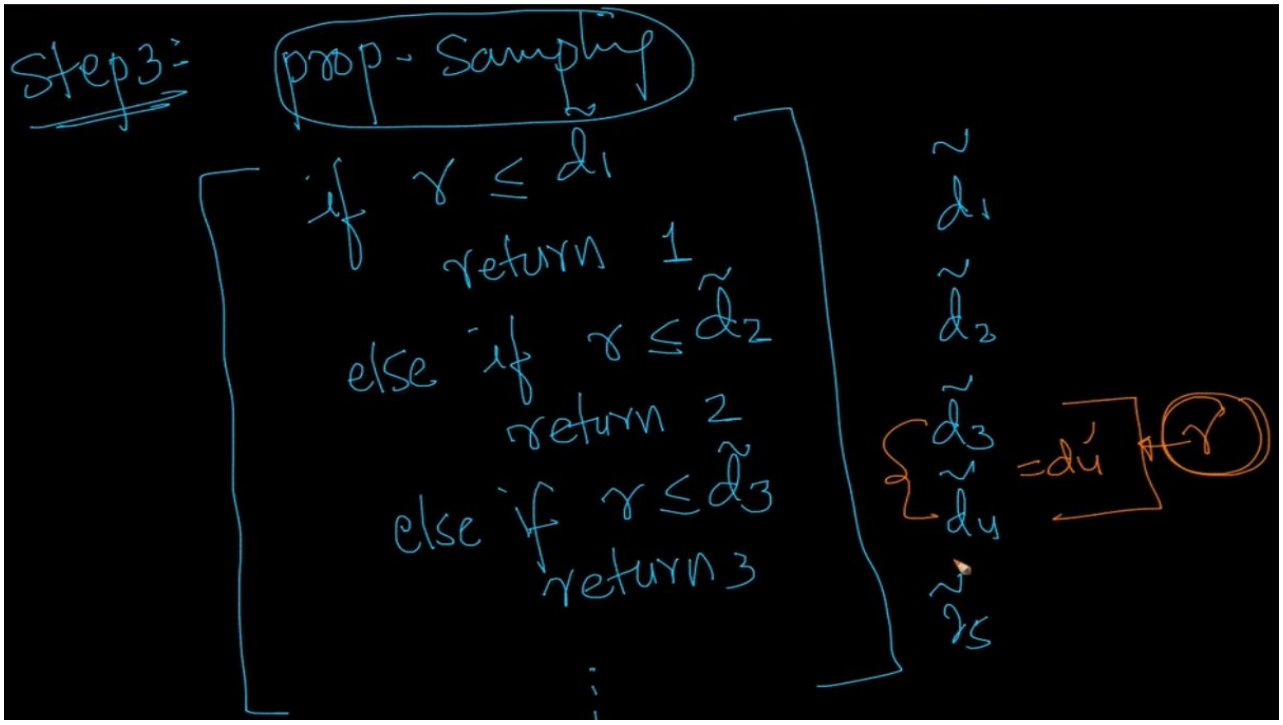


$d_i' = \frac{d_i}{S}$ .All the values between 0 and 1 and total sum is 1

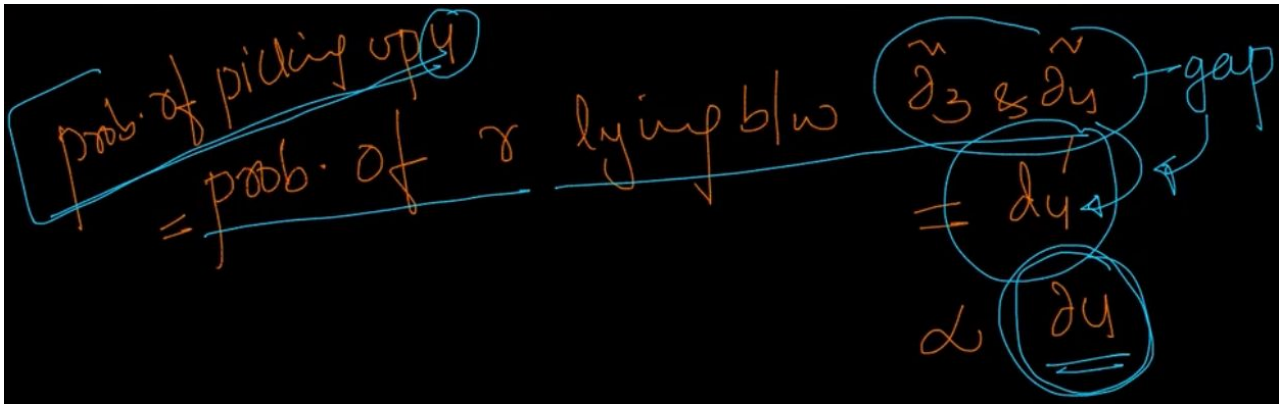


Here each value is calculated with the previous cumulative value and the current value as seen.





We've attached an if-else condition here . Here, r = 0.6 so it'll return 5 as $\overline{d_5}$ =1 . So it'll work and that's what we wanted because we wanted to pick d_5 *itself* . Let's make logic that prob of 'r' lying between $\overline{d_3}$ and $\overline{d_4}$ is d_4' (normalized value)



And d_4' α d_4 therefore d_5 will be returned because that's what we wanted