Linear Algebra

Sometimes we need to index a set of elements of a vector. In this case, we define a set containing the indices and write the set as a subscript. For example, to access x_1, x_3 and x_6, we define the set $S = \{1, 3, 6\}$ and write $\boldsymbol{x}_S$. We use the $-$ sign to index the complement of a set. For example $\boldsymbol{x}_{-1}$ is the vector containing all elements of $\boldsymbol{x}$ except for x_1, and $\boldsymbol{x}_{-S}$ is the vector containing all elements of $\boldsymbol{x}$ except for x_1, x_3 and x_6.

Matrix m*n ,
M = no of rows
N = now of columns

Probability :

Distribution :

For instance, if the random variable X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 for X = heads, and 0.5 for X = tails (assuming the coin is fair).

A **discrete probability distribution** (applicable to the scenarios where the set of possible outcomes is discrete, such as a coin toss or a roll of dice) can be encoded by a discrete list of the probabilities of the outcomes, known as a probability mass function.

On the other hand, a **continuous probability distribution** (applicable to the scenarios where the set of possible outcomes can take on values in a continuous range (e.g. real numbers), such as the temperature on a given day) is typically described by probability density functions (with the probability of any individual outcome actually being 0).

Types of Distributions :

1. Continuous
2. **Discrete**

What is Continuous variable :

The variable which never end by count called continuous variable .

Example . Can you count no of star in sky .

Lets take another example :

Continuous Variables would (literally) take forever to count. In fact, you would get to ‘‘forever’’ and never finish counting them. For example, take age. You can’t count ‘‘age’’. **Why not?** Because it would literally take forever. For example, you could be:
25 years, 10 months, 2 days, 5 hours, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picosends...and so on.

You *could* turn age into a discrete variable and then you could count it. For example:

- A person’s age in years.

Discrete Variable?

Discrete variables are countable in a finite amount of time. For example, you can count the change in your pocket. You can count the money in your bank account.

What is a Discrete Probability Distribution?

A discrete probability distribution is made up of discrete variables. Specifically, if a random variable is discrete, then it will have a discrete probability distribution.

For game 1, you could roll a 1,2,3,4,5, or 6. All of the die rolls have an equal chance of being rolled (one out of six, or 1/6). This gives you a **discrete probability distribution** of:

Roll	1	2	3	4	5	6
Odds	1/6	1/6	1/6	1/6	1/6	1/6

So here are limited chance of probability only 6 that's why it is discrete probability function .
For the guess the weight game, you could guess that the mean weighs 150 lbs. Or 210 pounds. Or 185.5 pounds. Or any fraction of a pound (172.566 pounds). Even if you stick to, say, betwen 150 and 200 pounds, the possibilities are endless:

- 160.1 lbs.

- 160.11 lbs.
- 160.111 lbs.
- 160.1111 lbs.
- 160.111111 lbs.

In reality, you probably wouldn't guess 160.111111 lbs...that seems a little ridiculous. But it doesn't change the fact that you *could* (if you wanted to), so that's why it's a **continuous probability distribution**.

Gaussian Distribution :

In probability theory, the **normal** (or **Gaussian** or **Gauss** or **Laplace–Gauss**) **distribution** is a very common continuous probability distribution.

Explain in Debug notes .

1 . Sampling Distribution means taking sample from main data and take mean of them .

3. Central Limit Therm states that as no of sample size increases lets say n = 30 means in each sample data point > 30 then data will be normally dist .
4. Mean of sample will be same of main data mean .

QQ plots are used to tell 2 data are from same dist or not .

They also used to check whether data belongs to normal dist or not .

We just simply sort all data . then find percentile values for x,y and then plot them on a graph . if all the points fit to straight line then its normal dist .

CLT says that whatever dist of data if we take more no of sample and mean of that sample . Then take mean of sample mean will give us normal dist .

CLT :

The example below performs this experiment and plots the resulting distribution of sample means.

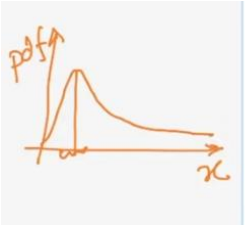
```
# demonstration of the central limit theorem
from numpy.random import seed
from numpy.random import randint
from numpy import mean
from matplotlib import pyplot
# seed the random number generator
seed(1)
# calculate the mean of 50 dice rolls 1000 times
means = [mean(randint(1, 7, 50)) for _ in range(1000)]
# plot the distribution of sample means
pyplot.hist(means)
pyplot.show()
```

Running the example creates a histogram plot of the sample means.

We can tell from the shape of the distribution that the distribution is Gaussian. It's interesting to note the amount of error in the sample mean that we can see in 1,000 trials of 50 dice rolls.

Further, the central limit theorem also states that as the size of each sample, in this case 50, is increased, then the better the sample means will approximate a Gaussian distribution.

Log Normal Distribution :



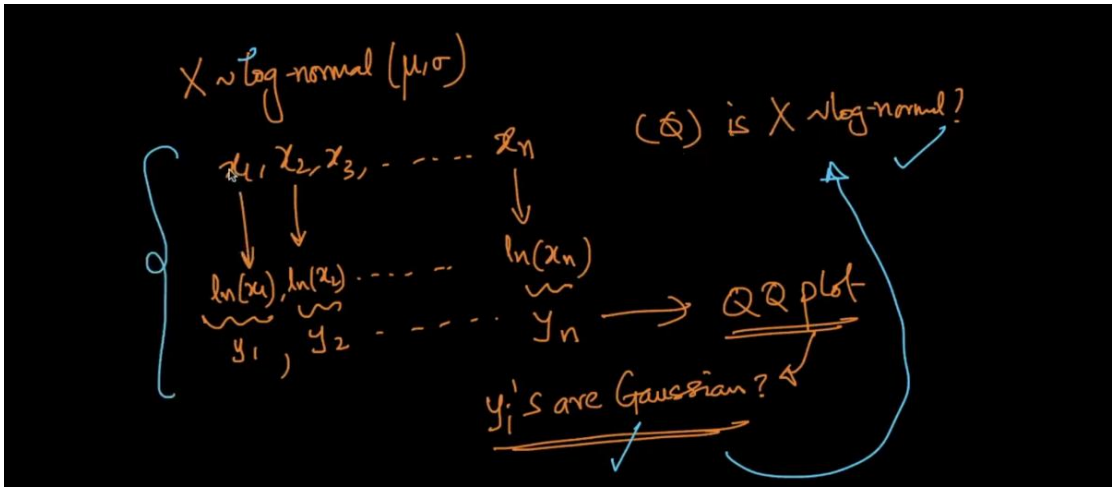
Examples include the following:

- 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

PDF says that lots of comments length are small that's why first part is at high peak and very few comments have large length .

How to check X is log normal or not ?



Lets say we have X = {x1,x2 …….. x n }

Then we create new variable lets say

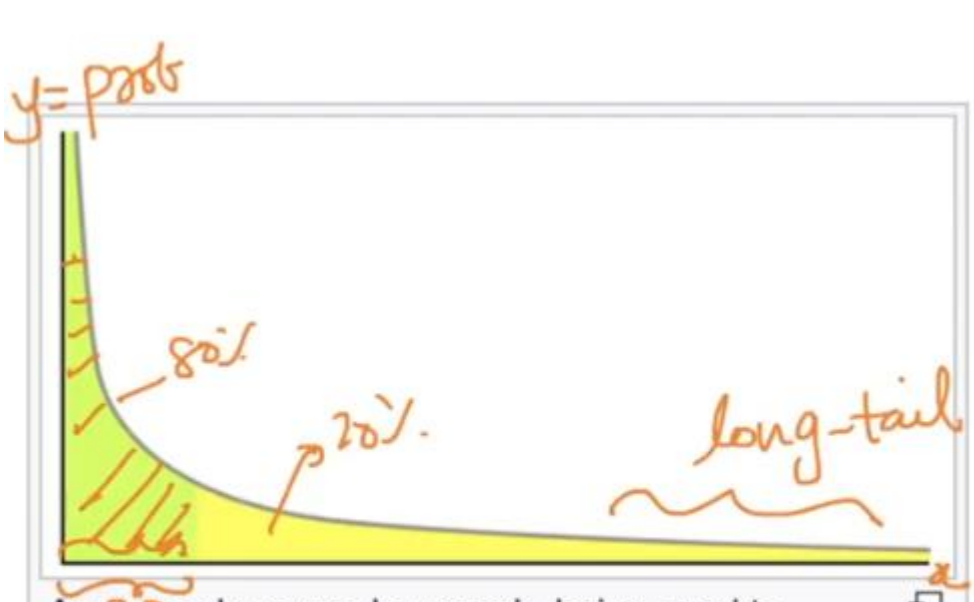
Y = {log(x1),log(x2),log(x3) …… log(x n)}

Now we know QQ plot use it on X and Y and see is Y is Gaussian or not If Y is Gaussian then we can say X is log normal .

This dist is widely used in finance section .

Log dist occurs many many times .

Power Law Distribution :

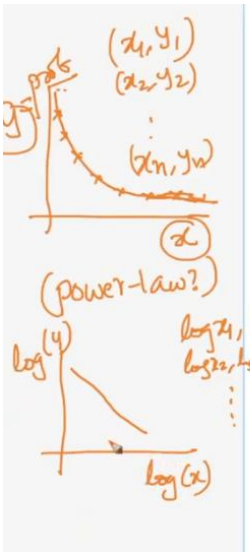


An example power-law graph, being used to demonstrate ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80–20 rule).

This distribution follows 80-20 Rule.
Means 80% of data we can wind in 20% interval means that green part in image .

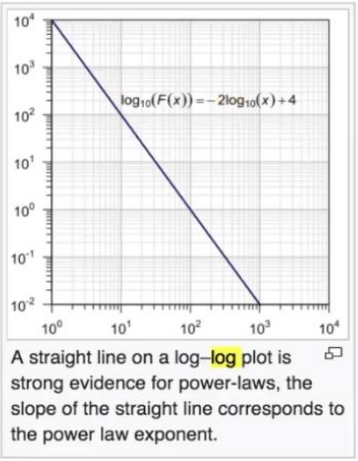
Whenever dist follows power law its called Pareto dist .

How can we say dist is power law or not ?



produces plots that are difficult to interpret; for this reason, s… called Hill horror plots [30]

Log–log plots are an alternative way of graphically examining the tail of a distribution using a random sample. This method consists of plotting the logarithm of an estimator of the probability that a particular number of the distribution occurs versus the logarithm of that particular number. Usually, this estimator is the proportion of times that the number occurs in the data set. If the points in the plot tend to "converge" to a straight line for large numbers in the x axis, then the researcher concludes that the distribution has a power-law tail. Examples of the application of these types of plot have

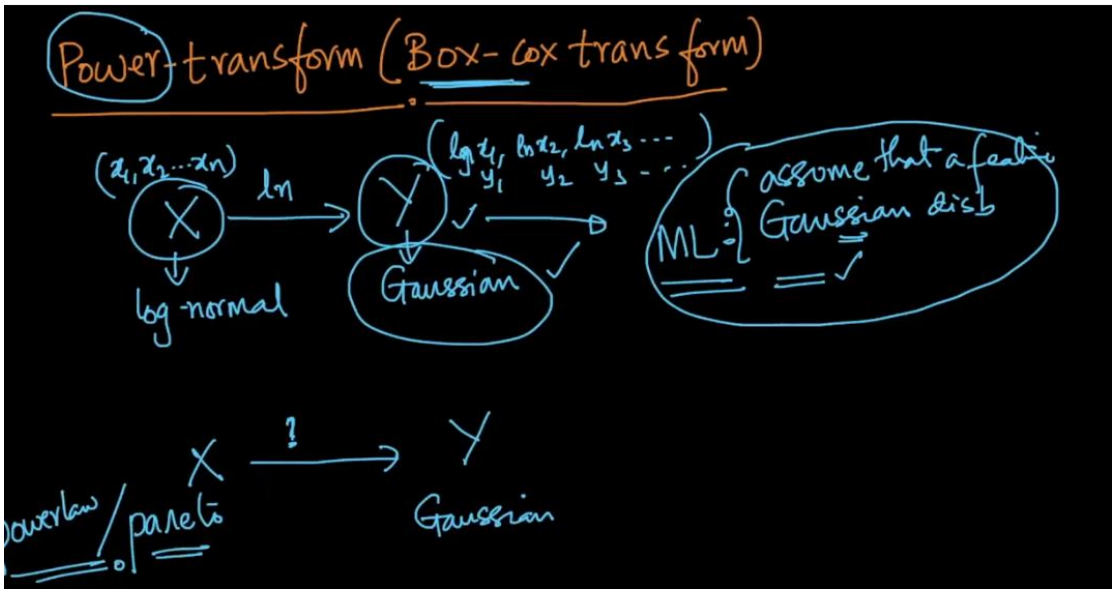


A straight line on a log–log plot is strong evidence for power-laws, the slope of the straight line corresponds to the power law exponent.

Lets say we X data and Y probability graph .

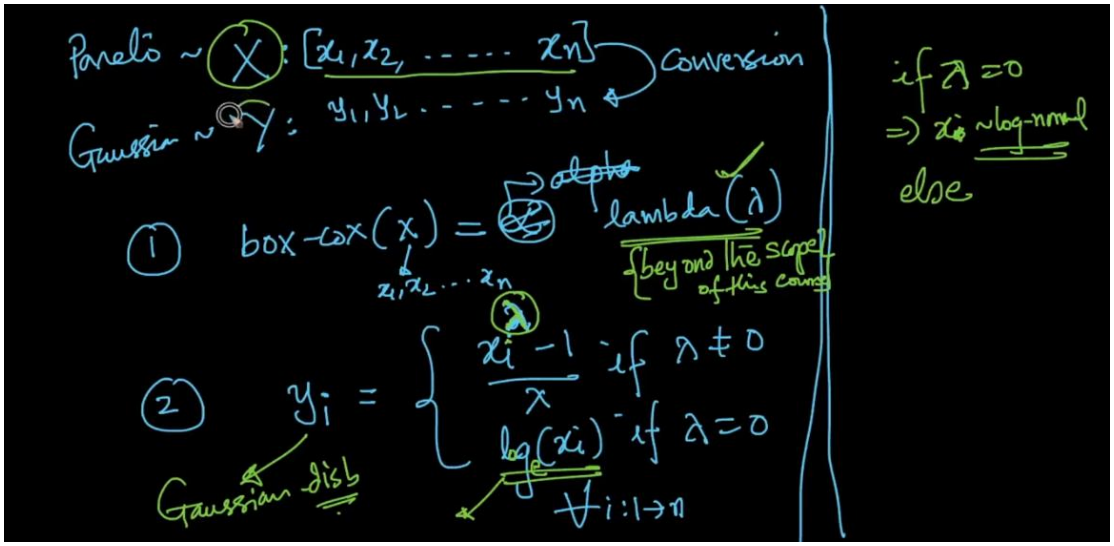
So now we will take log of both of them and after that we will plot and if we get straight line then we can say it follows power law dist .

Box - Cox Transform :



Now as we see above Pareto dist and power law dist occurs a lot so is there any way we can convert this into Gaussian because from Gaussian we can make lot of different analysis .

Answer is yes called as Box - Cox transform .



Now lets say we have data X which is Pareto dist and now we want to convert to Gaussian so what we can do is convert X to Y .

Step 1 : Call python box-cox function and pass X as a argument . (Return lambda)

Step 2 : Calculate Y using above formula .
If in case lambda = 0 it means that X(i) value is log normal so just take natural log of that value so we get normal dist as we study in log normal dist .

Application of Non Gaussian Dist :

Now we have example of Dam .

Lets say civil engineers are designing a new dam so they need to consider lots of factors before construction ex . height of dam,maximum rainfall per day etc ..

So lets say they have 200 data points means 200 data about rainfall .

So now we want to know P(rain>20cm) so to solve this problem we have 2 methods .

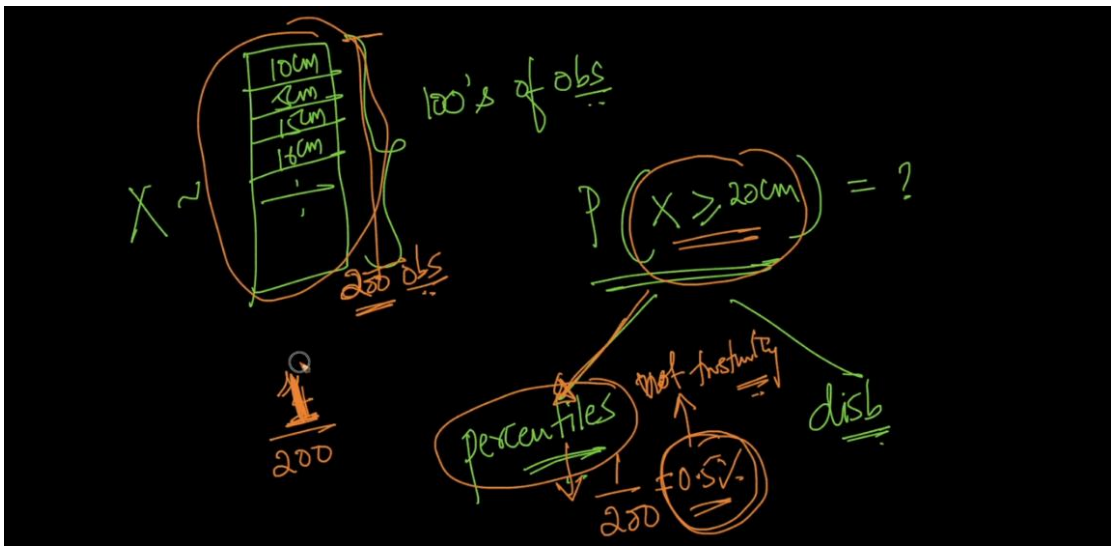
1 : Percentile

2 : Distribution

Now lets say we use percentile so in we check our data and we find only one observation which has value > 20 .

So percentile will = 1/200 = 0.5%

So from just one observation we can not tell anything about probability so that's why Percentile will not work here .



So we use distribution for all data and then from PDF and CDF we can give max answer of questions this is power of distribution .

Co-variance :
It is used to measure how much 2 random variable vary together .

But as scale change Co-variance also change for same data so this is very big dis of Co-variance .

To remove this we use Pearson correlation c (PCC) which is calculate simply dividing COV(X,Y) with Standard deviation of X and Y .

It is always lies between -1 to + 1 .

Bootstrapping Method :

Now in this method we don't know dist but now I want to find 95% confidence interval for my data how can we solve this ?

Lets say I have data X .

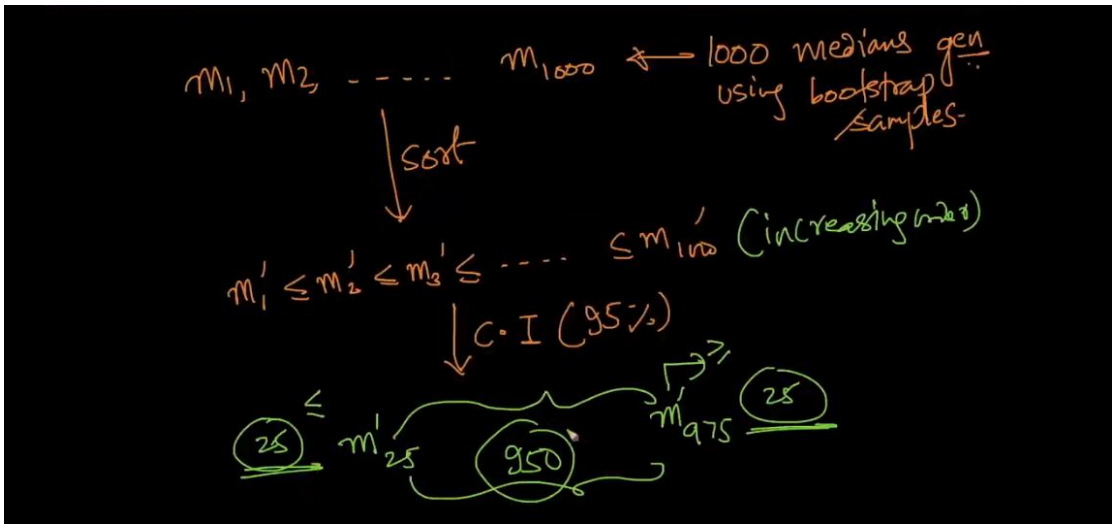
From X I will create small sample S (random sample)

From S we crate lets say 1000 samples .

Now we will find median of 1000 sample and will sort them ascending order .

After that now we have sorted 1000 median .
So 95 % means

We will take range between from 25th median to 975th median so we will get 950 values = 95% .



Empirical bootstrap based Confidence Interval

```
In [3]: import numpy
        from pandas import read_csv
        from sklearn.utils import resample
        from sklearn.metrics import accuracy_score
        from matplotlib import pyplot

        # load dataset
        x = numpy.array([180,162,158,172,168,150,171,183,165,176])

        # configure bootstrap
        n_iterations = 1000
        n_size = int(len(x))
```

```
# run bootstrap
medians = list()
for i in range(n_iterations):
    # prepare train and test sets
    s = resample(x, n_samples=n_size);
    m = numpy.median(s);
    #print(m)
    medians.append(m)

# plot scores
pyplot.hist(medians)
pyplot.show()

# confidence intervals
alpha = 0.95
p = ((1.0-alpha)/2.0) * 100
lower =  numpy.percentile(medians, p)

p = (alpha+((1.0-alpha)/2.0)) * 100
upper =  numpy.percentile(medians, p)
```

```
upper =  numpy.percentile(medians, p)
print('%.1f confidence interval %.1f and %.1f' % (alpha*100, lower, upper))
```

Null Hypothesis :

The null hypothesis can be thought of as a *nullifiable* hypothesis. That means you can nullify it, or reject it. What happens if you reject

the null hypothesis? It gets replaced with the alternate hypothesis, which is what you think might actually be true about a situation. For example, let's say you think that a certain drug might be responsible for a spate of recent heart attacks. The drug company thinks the drug is safe. The null hypothesis is always the accepted hypothesis; in this example, the drug is on the market, people are using it, and it's generally accepted to be safe. Therefore, the null hypothesis is that the drug is safe. The alternate hypothesis — the one you want to replace the null hypothesis, is that the drug *isn't* safe. Rejecting the null hypothesis in this case means that you will have to prove that the drug is not safe.

In our height example if we get p value lets say 20% means we don't have sufficient evident to say that height between 2 cities is different so we accept null hypo that means we OK with null hypo which is height between 2 city same .

If p value 3% means we have 97% of data which is saying that they found difference in 2 city height

that means we enough evidence 97% to say that yes null hypo is wrong so reject it and accept alternative hypo .

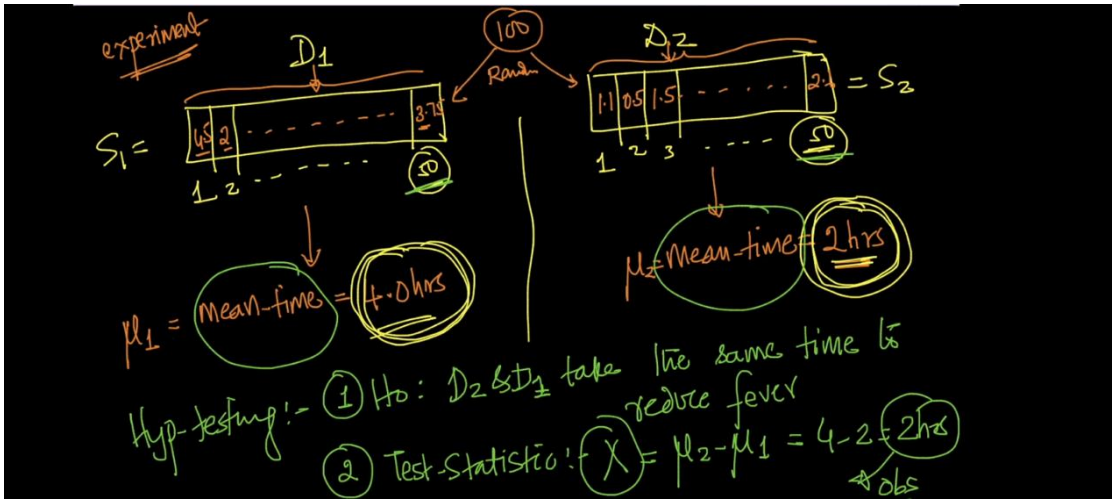
https://www.statisticshowto.datasciencecentral.com/support-or-reject-null-hypothesis/#noP

Use above link for more detail .

Null hypo example :

Lets say we have 2 medicines m1 and m2 and claim is m2 is more powerful than m1 and reduce fever in half time of m1 .

So we want to hypo test to check whether that new claim is correct or wrong .

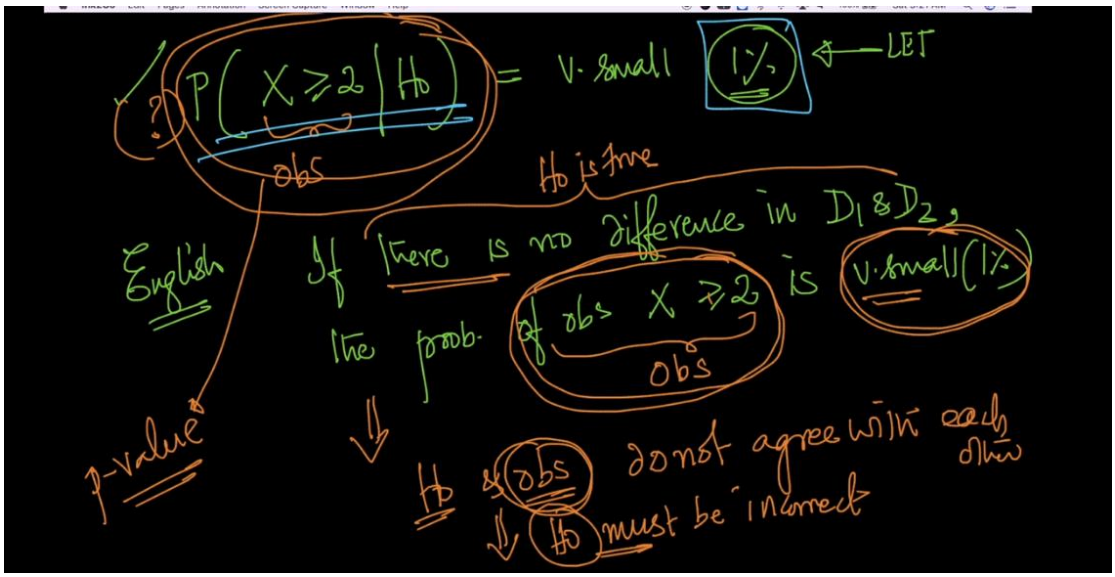


Now lets say we have 100 patients we split randomly in 50-50 group . 1 group treat by m1 and other by m2 .

Then we find mean of two group .

Our hypo is there is no diff in medicine .

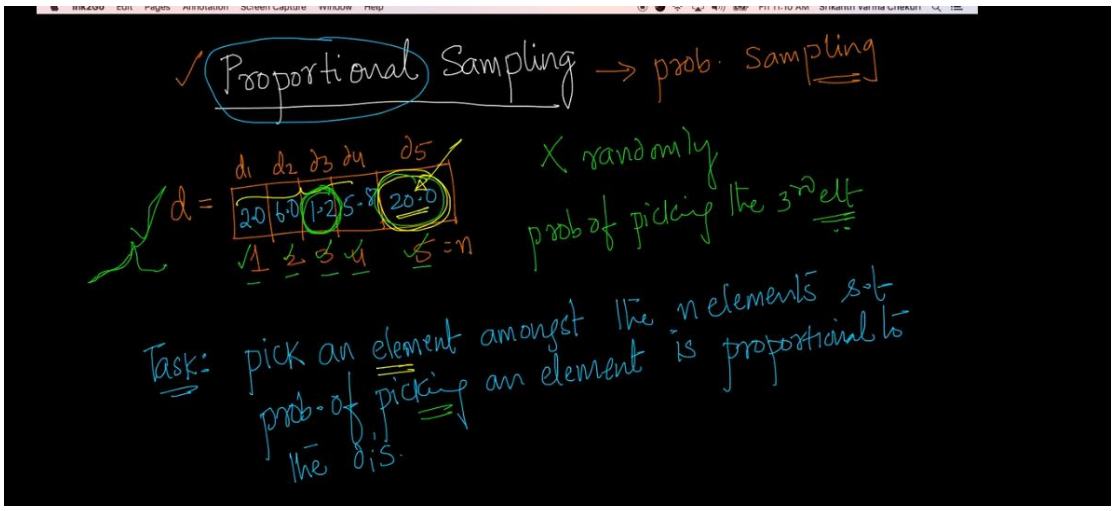
See above image .



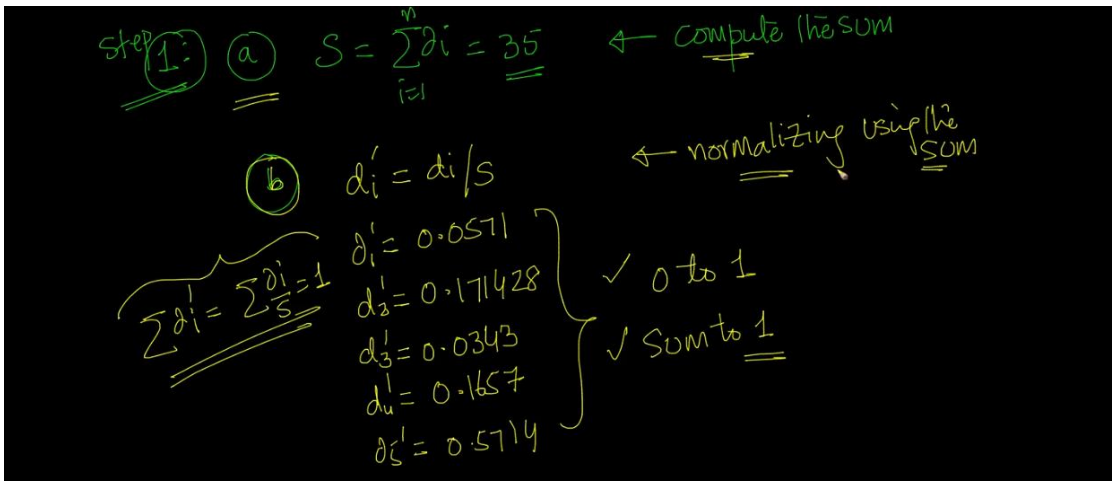
Now see our p value = 1% what it means ?

It means there is only 1% probability null hypo and observation is correct . so this is very very less probability so we reject it and we say m2 is more powerful than m1 .

That means we have 99% probability and evidence to proof that null hypo Is wrong so we reject it .

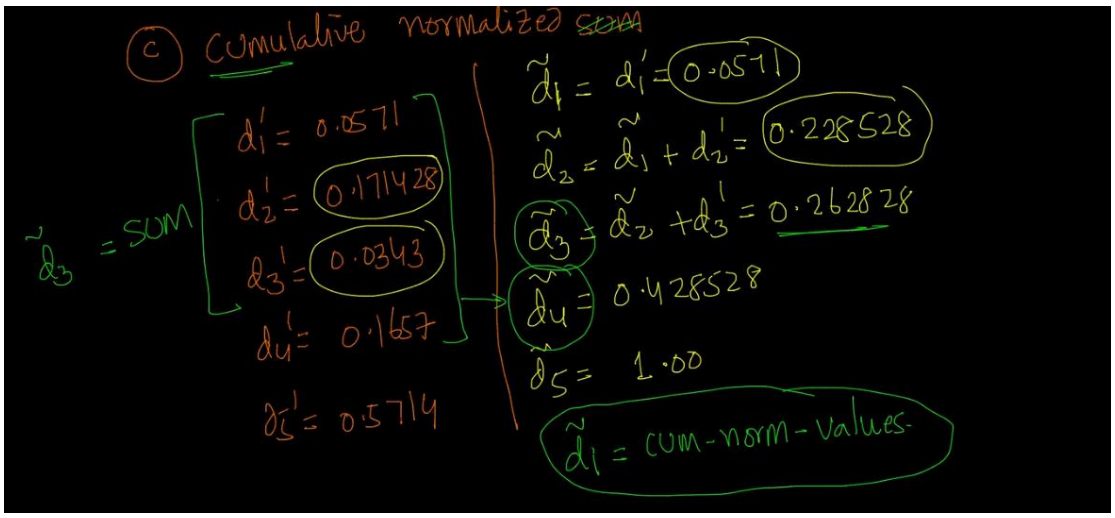


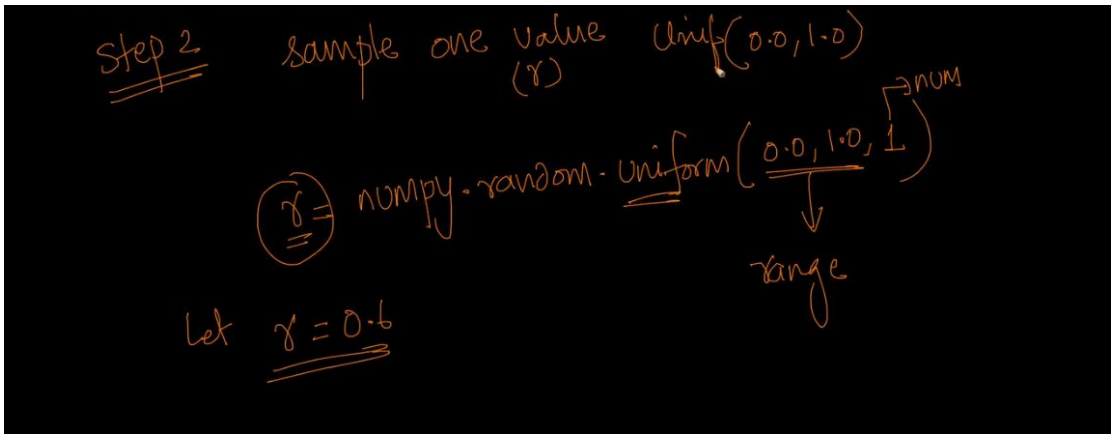
Here we want to make more probability of picking last value because it is largest value .



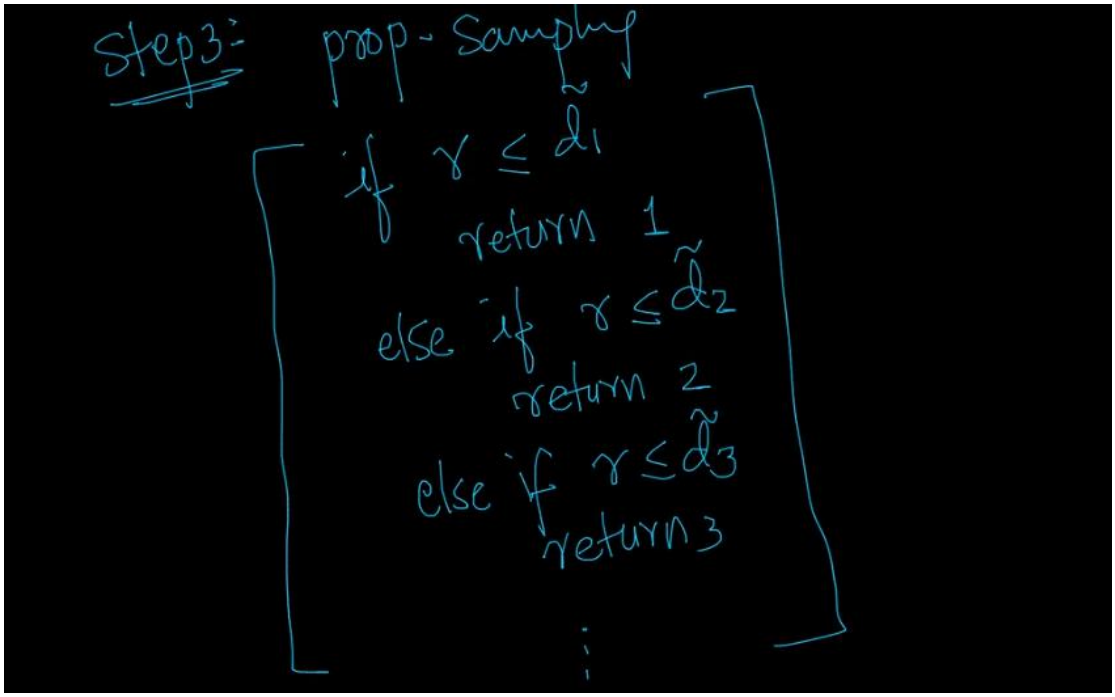
1 . Find sum of all points .
2. Find new d' which is = each data point / sum

3. So by doing this we get normalization so all values will lies between 0 and 1 and sum of all points will be 1 .





Now step 3 we will perform Proportional Sampling .



So now our r =0.6 we will use above if condition and as per that it will return 5 .

So we take value at 5th position that is equal to 20 in this way we achieved our target .