6/Oct

Page No.
Date

Statistics is the science of collecting, organizing and Analyzing data.

Data :- facts or piece of information.
Eg :- Height of students in a Class
140 cm, 145 cm, 146 cm, 180 cm, 171 cm.

Eg :- Gender of person visiting a doctor (8 am - 4 pm)
(M, M, F, T, F, M)

Statistics is of Two types → Descriptive Stats.
→ Inferential Stats

Descriptive Stats
↳ It consist of organizing and summarizing of data.

Inferential Stats
↳ It consist of using data you have measure to form conclusion.

D-S → measure of central tendency (Mean, Median Mode)
→ measure of dispersion
↳ variance
↳ Standard Deviation.
→ Different types of Distribution of Data
→ Histogram, pdf, pmf.

I.S → Diff of Type of Test

- ↳ Z Test
- ↳ t-test
- ↳ CHI square test
- ↳ ANOVA test.

} Hypothesis testing. H_0, H_1 p-value, Significance value.

Eg: Let consider there are 20 classroom in a college and let say you have collected the Ages of student in one class.

→ Ages of Class : {19, 21, 18, 34, 22, 21, 25, 24}

Descriptive Stats Questions.

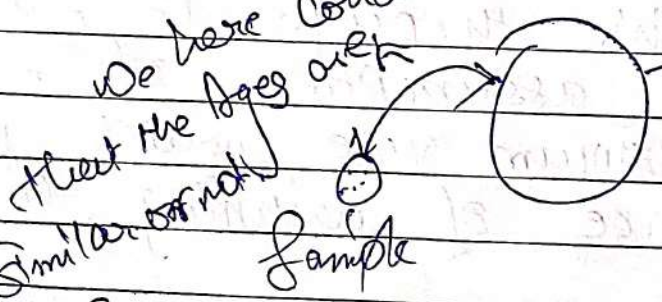
↳ What is the common Age in your Stats class.

Ans → Mean (Ages)

↳ which is a part of Central tendency.

Inferential Stats Question.

↳ Are the Ages of student in the classroom similar to ~~total~~ ~~you expect in~~ the Ages of the student in the University.



In I.S, we have that Sample & Population is going to be performed.

Page No.
Date

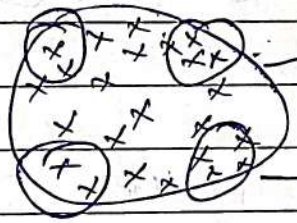
Tut 2 Population & Sample, And Sampling Technique.

Population:- The group that we are Interested in studying.

Sample :- It is a subset of a population.

Eg:- EXIT POLL ⇒ Assumption made on the basis of sample data, that what are the chance of winning the state election for ~~the~~ which party?

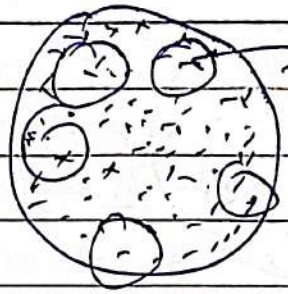
STATE A — Age > 18 year



So, ~~On the Basis~~ By reaching out to the sample data and getting the insights through it, we'll make an assumption on the basis of maximum vote which party has the chance of winning the election.

But the EXIT POLL, it is not always True. This is just an assumption.

STATE B ka Average height of the people.



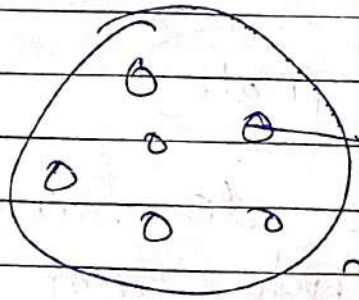
Note → As we choose sample from the population, But what are conditions/methods or process of taking the samples from the population. what is the criteria behind it, So to know that we should know Sampling Techniques

Sampling Techniques ⇒ The main motive of this is, that when we choose sample so on what basis we are choosing it.

→ The goal of sampling is to create a sample that is representative of the entire population.

Population is denoted by (N)

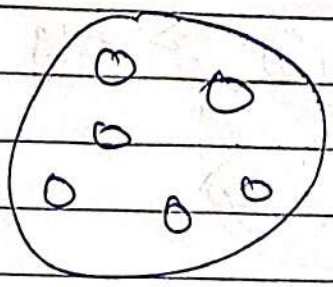
Sample is denoted by (n)



Lets, Find Out how many types of sampling techniques are there.

Types of Sampling.

① Simple Random Sampling.



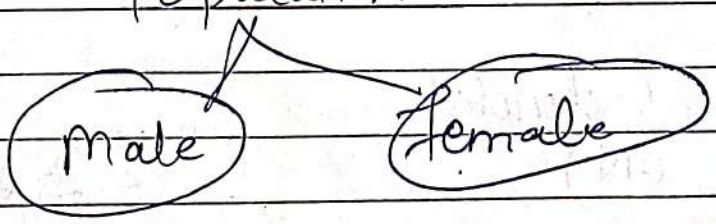
Randomly we are taking any samples without any condition applying to it.

When performing Simple Random Sampling, every member of the population (N) has an equal chance of being selected for your Sample (n)

② Stratified Sampling:

Stratified means layering, we also define it as ~~State~~ Strata. And these layering are non-overlapping groups.

Eg:-



If I wanna do survey on male, that I need info from male gender, Job I'll recruit to Male.

Some goes for female.

Aise he, you multiple groups, but make sure that these groups shouldn't

③ Systematic Sampling :-

Assume, we are doing a survey outside a Mall every 4th person, I'm asking ~~for~~ to fill the needful for the survey.

Whenever you decide for (select) samples, you decide a term "4" (nth) and whenever I see the nth person, I'll approach ~~to~~ that person for the survey.

④ Convenience Sampling :- If we want to do Sampling., expertise or there is term called "Voluntary Response Sampling"

This term is very important with Convienience Sampling.

Eg :- I want to do a survey on Data Science.

Data Science

Will Approach to them, those who have Knowledge in DS, experts, Student in DS.

We can't approach Doctor. Right!

Exercise !
→ EXIT POLL {Random}
→ Disease Information (Convenient Sampling)
→ Household expenses (Strategic Sampling)
Main :- (Women)
Target

TOT 3 What are Variable & It's Types.

Defn: Variable is a property that can take on multiple / many values.

Eg:- age = 12, 13, 40 Height = 172 cm, 172.5 cm, 180 cm

Weight = 72 kg, 72.5 kg, 73 kg.

Variable → Singular mode.
Whenever we talk about variable we talk with respect to only with a single value.

Ages = [12, 24, 48, 60, 100]
↳ plural mode.

* Types of Variable

Quantitative Variable

Discrete Variable (whole no.) VS Continuous Variable

Discrete Variable: The value where Discrete variable will be assigned will always be a whole Number
Eg:- no. of Bank Acc, no. of children

Continuous Variable: It takes Infinite Values.
Eg:- Height, Weight,

Qualitative Variable (Categorical)

Based on the properties this variable is further classified as.
Eg:- Gender (Male, Female)
↳ It is divided on some properties.

Types of Flower
(Rose
(Lily
(Cactus.

(c) Median
(B) Mode

Defi:- Central Tendency refers to the measure used to determine the "center" of the Distribution of data.

And (mean, Median, Mode) are the techniques to find out the center of Distribution of data.

Eg:- {1, 2, 2, 3, 4, 5} lets find out Mean (Average)

Usually when we talks about (mean, median, mode) we mostly focus on two types of Data → population Data (N) μ → Population Mean
→ Sample Data. (n) $\bar{x}$ → Sample Mean

population Data (N)	Sample Data (n)
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

→ Population Data

$X = \{1, 2, 2, 3, 4, 5\}$

let find the Average (Mean)

② Median ?

~~$X = \{1, 2, 2, 3, 4, 5\}$~~ ~~$X = \{12$~~

~~$\mu = 2.83$~~

find.

② Median

$X = \{1, 2, 2, 3, 4, 5, 100\}$ → This number is called outliers, which is out of scope.

~~$\mu = 2.83$~~

$\mu = 2.83$ (previous Mean)

$\mu = \frac{1+2+2+3+4+5+100}{7}$

$\mu = 16.71$

There is a huge diff in both of this μ (Mean). the value 16.71 is out of scope just because 100 got added, which is a case of outliers, so, to get rid of these outliers we'll use median.

Median says ✳ There are 3 steps.

① Sort all the numbers

② Find the Central element ← odd length / even length

Odd length → $\{1, 2, 2, (3), 4, 5, 100\}$ → median = 3

Because there are 7 elements

Even length/observation = $\{1, 2, 2, 3, 4, 5, 100, 101\}$

$= \frac{3+4}{2} = 3.5$

(B) Mode :- Most Frequent Elements.
(It check the frequency of the data element of String data / Numerical data.

{1, 2, 2, 3, 3, 3, 4, 4, S, S, S, S}

most frequent Element.

We use Mode, when we have a categorical feature

Eg :-

	Age	Weight	Gender
Missing value will get replace by the Mean	24	-	M
	25	-	F
	-	78	-
	-	80	-
	26	82	
	27	-	

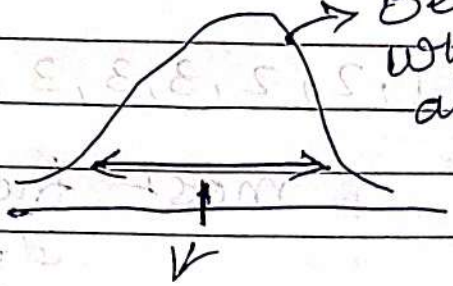
Suppose Here Male is repeated Value, then ~~H to~~ missing value will get replace by M - Male.

Weight, 9] Incase there is Outlier will use Median Based on Even or odd observation

T.S Measure of Dispersion - Variance & Standard Deviation.

(Spread)

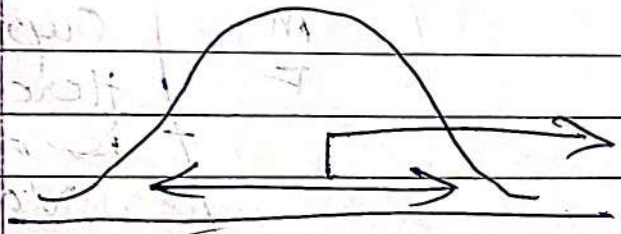
→ ① Variance
→ ② Standard deviation



Bell curve which is also called as Normal Distribution

whenever we speak about this central element, it can be either our mean, median, mode.

↳ And it specifies measure of central tendency



This spread, which is also called as Dispersion

And this spread is determine using two important things → Variance
↳ Standard Deviation

population (N)	Sample (n)
$\mu = \frac{\sum_{i=1}^{N} (x_i)}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} (x_i)}{n}$
Population variance → Sigma $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	Sample Variance → "s" square $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$ ← Bessel Correction
	Degree of freedom.

Eg: $X = \{1, 2, 2, 3, 4, 5\}$

x	$\bar{x}$	$(x - \bar{x})$	$(x - \bar{x})^2$
1	2.83	−1.83	3.34
2	2.83	−0.83	0.6889
2	2.83	−0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
			10.84

$\bar{x} = 2.83$

(Sample Variance) $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

$= \frac{10.84}{6-1}$

$= \frac{10.84}{5}$

Sample Variance: $s^2 = 2.168$ → (Spread of the Data, Dispersion we can say)

Standard Deviation formula →

↳ $s.d = \sqrt{\text{variance}}$

$= \sqrt{2.168}$

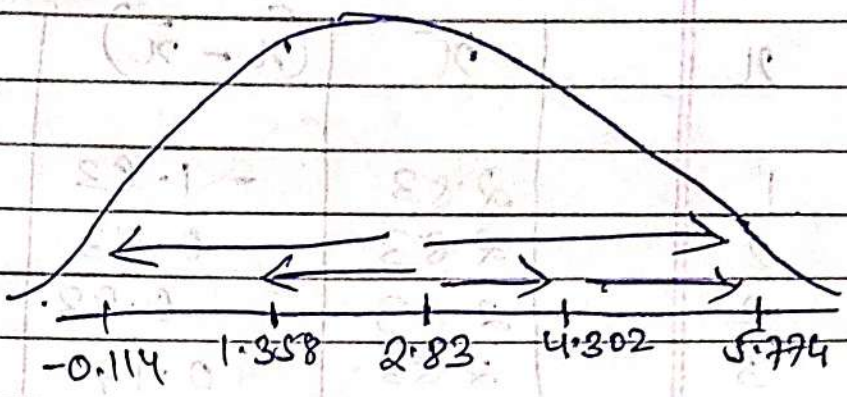
$= 1.472$ → Sample Standard Deviation

Page No.
Date

Dataset Insights

Sample Mean $\bar{x} = 2.83$,

Sample S.D = 1.472



$2.83 + 1.472 = 4.302$

$4.302 + 1.472 = 5.774$

$2.83 - 1.472 = 1.358$

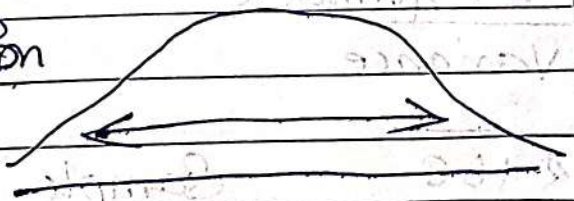
$1.358 - 1.472 = -0.114$

Here you got the idea of Standard Deviation But still you dont know what is the work or Role of Variance.

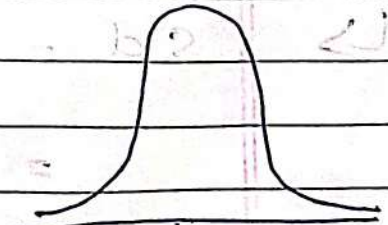
Variance decides the entire Spread or dispersion.

Lets say → Variance — Big no. S.D will also be Big no.

Spread/Dispersion will be more



Variance - Small no. S.D will also be Small no.



Spreadness will be less, But height will be more.

Note → ① If Variance is a big number Spread is also increasing.

② If Variance is small number spread is also decreasing.

T.6 Percentile and Quartiles

Diff betn Percentage & Percentile

Percentage = {1, 2, 3, 4, 5}

% of the numbers that are even?

% of even = $\frac{2}{5} \times 100 = 40\%$

So this was percentage

Now, Percentile :- A percentile is a value below which a certain percentage of observations lie.

eg :- 95 percentile means that the person has got better marks than 95% of the entire student.

Dataset :- 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, [10], 11, 11, 12

What is the percentile ranking of 10?

Step 1 :- Note how many numbers are there in a dataset :- $x = 10$

$n = 20$

Percentile Rank of 10 = $\frac{\text{total no. of value below } x}{\text{total no. of observation } (n)} \times 100$

$= \frac{8}{20} \times 100$

$= 80$ percentile.

Now, what does this 80 percentile indicates?

The value 10 is 80% bigger value than entire distribution value.

Now for 12, find the percentile value for 12.

$x = 12$

$n = 20$

Percentile Rank of 20 = $\frac{\text{total no. of Obs below } x}{n} \times 100$

$= \frac{19}{20} \times 100$

$= 95$ percentile

The value 12 is 95% bigger value than entire distribution.

$$\begin{array}{r|l} & 9.5 \\ 2 & \overline{)19} \\ & \underline{-18} \\ & 10 \\ & \underline{-10} \\ & 0 \end{array}$$

(2) Different Question :-
What value exists at percentile Rank of 25. (what value exist at 25 percentile)

Formula :-

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{25}{100} \times (20+1)$$

$$= 0.25 \times 21$$

$$= 5.25$$

5.25 is not the value in the Index for the given dataset

Index = 5.25

Given dataset → 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, ----

0 1 2 3 4 5.25 6

will Calculate Avg of 5 & 6 element which is nothing but 5, 5

$$\frac{5+5}{2} = \frac{10}{2} = 5$$

So, 5 is the value, which lies at 25 percentile.

Quartile

Box Plot

→ 1st Quartile (Q1) – 25%

Median → 2nd Quartile (Q2) – 50%

→ 3rd Quartile (Q3) – 75%

Inter Quartile Range → IQR = Q3 – Q1

→ Lower Bound (LB) = Q1 – 1.5 (IQR)

→ Upper Bound (UB) = Q3 + 1.5 (IQR)

Quartile is

To find Outlier, we Specifically use Box plot,

To know how Box plot is created, we have to use Quartile

IQR (Inter Quartile Range) → And from this IQR we find the Outliers.

T.T How to construct a Box Plot for Outliers.

But Before that we'll solve a problem. (Chaitanya Sir)

Dataset ⇒ {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 9, 9, 27}
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

First will find Q1 - 25% percentile

formula ⇒

$$Q1 = \frac{\text{percentile}}{100} \times (n+1)$$

$$\frac{25}{100} \times (19+1)$$

$$\frac{25}{100} \times 20$$

$$\frac{50}{10}$$

= 5th Index, Value = 3

Now, Q2 - 80% percentile

$$Q2 = \frac{\text{percentile}}{100} \times (n+1)$$

$$= \frac{50}{100} \times 20$$

Q2 = 10th Index, Value =

Now Q3 — 75% percentile

$$Q3 = \frac{\text{percentile}}{100} \times (n+1)$$

$$= \frac{75}{100} \times (20)$$

$$= 15\text{th index, value} = 7$$

Now IQR (Inter Quartile Range)

$$IQR = Q3 - Q1$$

$$= 15 - 5 \qquad 7 - 3$$

$$IQR = 10 \qquad 4 \Rightarrow IQR$$

Now LB (Lower Bound)

$$LB = Q1 - 1.5(IQR)$$

$$= 5 - 1.5(10) = 3 - 1.5(4)$$

$$= 5 - 15 \qquad = 3 - 6$$

$$= -10 \qquad = 6 - 3$$

And UB (Upper Bound)

$$UB = Q3 + 1.5(IQR) \qquad 7 + 1.5(4)$$

$$= 15 + 1.5(10) \qquad 7 + 6$$

$$= 15 + 15 \qquad 13$$

$$= 30$$

$$UB = 13$$

for outliers:-

→ Five Nos Summary And Box Plot.

1. Minimum
2. First Quartile (25%.) Q1
3. Median
4. Third Quartile (75%.) Q3
5. Maximum

All these components are used to create Box plot.

And the main aim of this Box plot is that it helps to detect the outliers

Removing the Outliers.

This Data should be in sorted.

Dataset → {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

→ So, to detect the outlier, this technique is most favourable in Machine learning.

And,

[lower fence ⟷ Higher fence]

If lower fence is 5, or it can be any value. So below the lower fence value whatever number lies will be Outliers, And for Higher fence = 30 or any value, whatever value above higher fence value lies

Here, this 27 is a outlier, which is entirely different from all the Distribution value. But we cannot just confirm that it's an outlier, Because there would be lot of Data. Even there could be

IQR = Q3 - Q1

Page No.
Date

Lower fence = Q1 - 1.5 (IQR)

Higher fence = Q3 + 1.5 (IQR)

$$Q1 = \frac{\text{percentile}(25)}{100} \times (n+1)$$

$$Q3 = \frac{\text{percentile}(75)}{100} \times (n+1)$$

LB — (-3) ⟷ (13) — UB.

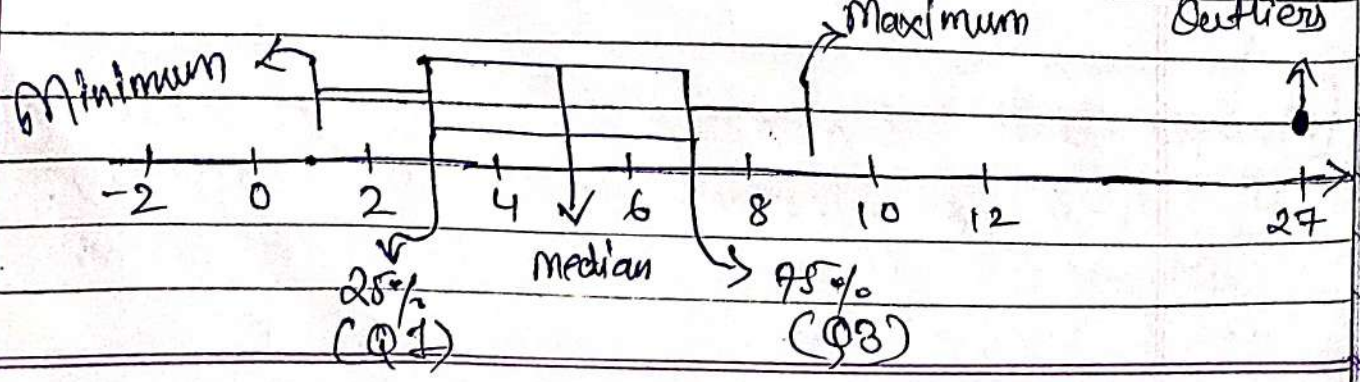
This is the range from -3 to 13

So, 27 K definitely are outliers. So we can remove it or either we can plot it with a Box plot.

So, let's move to the 5 components, as per the Dataset

1. Minimum value = 1
2. First Quartile (25%) Q1 = 3
3. Median = 5 (If we calculate, count of elements)
4. Third Quartile (75%) Q3 = 7
5. Maximum = 9 (we won't consider outliers)

Now, will create Box plot.



apply Z-score, And this process is called Standardization.

Z-score :- It tells how much standard deviation it is away from the mean.

Formula :-

$$\text{Z-score} = \frac{x_i - \mu}{\sigma}$$

$\sigma \rightarrow$ standard deviation.

X	μ	$x_i - \mu$	$(x_i - \mu)^2$	working.
1	2.83	-1.83	3.35	1+2+2+3+4+5
2	2.83	-0.83	0.68	
2	2.83	-0.83	0.68	
3	2.83	+.83	0.17	0.03
4	2.83	2.83	1.17	1.36
5	2.83	3.83	2.17	4.70
$\mu = 2.83$				10.80

Mean (μ) = 2.83

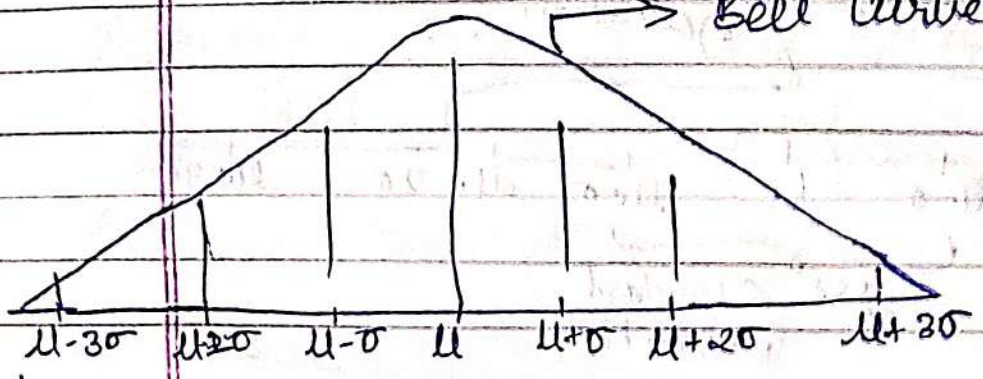
finding Variance

$\rightarrow$ variance formula

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

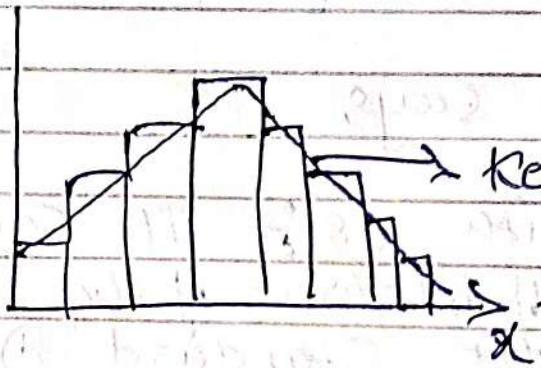
$$\sigma^2 = \frac{10.80}{6}$$

T.9 | Normal Distribution / Gaussian Distribution



To know it, in simple term that how a Bell Curve is made will take eg of histogram.

Eg: lets take an histogram,



the Bell curve is apply after smoothening process, and this smothening is called. kernel density estimator.

And this kernel density, converts this histogram into pdf (probability density function).

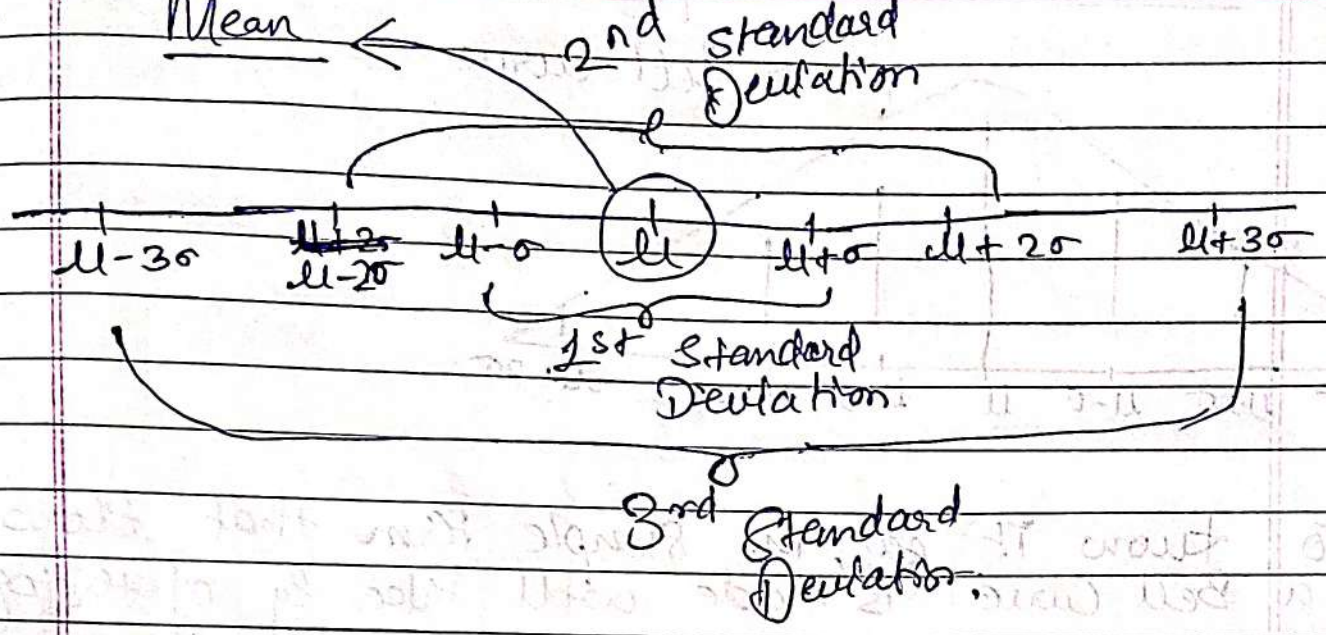
. And after doing this process, we get this Bell curve..

→ ① Symmetrical - [left area and Right area of the curve are both equal.]

for eg:- we have data set: {1, 2, 3 ... y}

* Symmetrical means 50% of number is on the left side of mean and remaining 50% is on Right side of the Mean.

(2) Mean



68 - 95 - 99.7 [Empirical Rule]

Empirical Rule says,

Lets go with 68:- It says that 68% of entire distribution of the data lies within 1st Standard Deviation.

Let assume, we have a data set of 100 elements. So, 68 data element will lie within 1st standard Deviation as per Empirical Rule.

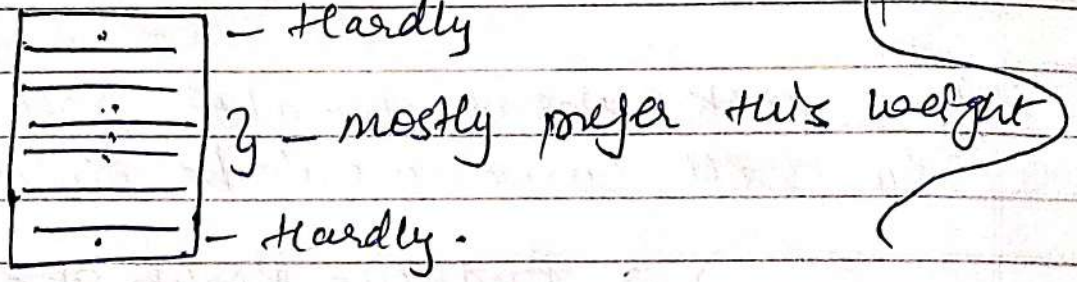
Now 95, 95% of the data of this Distribution ~~following~~ falling within 2nd Standard Deviation.

And 3rd Standard Deviation of Data of this Distribution covers 99.7% of data.
only if it follows Gaussian Distribution or Normal Distribution.

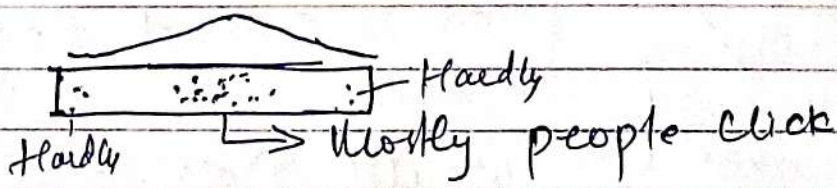
Note:- Whenever it is Gaussian Distribution or Normal Distribution then definitely it follows the 3 Sigma rule of Empirical which are [68-95-99.7]

let see some examples, like what kind of Dataset does normal or Gaussian Distribution follows :-

Eg:- Gym weight plates



Space bar



Note:-

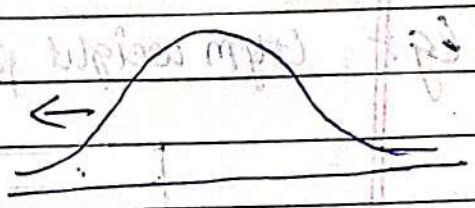
Note:- Whenever it is Gaussian Distribution or Normal Distribution then definitely it follows the 3-Sigma rule of Empirical rule [68-95-99.7] and this is an important property of Gaussian or Normal Distribution. and with the help of this you can solve many problem as

① you can find Outliers (By the 3rd standard Deviation, like whatever Data lies outside the 3rd standard Deviation will be count as Outliers.)

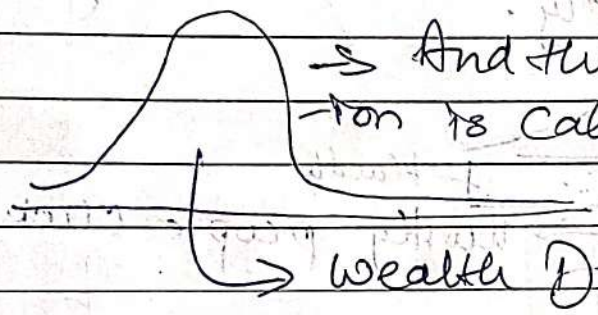
T.10 Central Limit theorem —

Weight Data (x) = {65, 72, 83, 55, 64, 67 - - - -}
Variable ← weight of a population.

So, the Bell curve will be Normal / Gaussian Distribution.



Q. If I talk about wealth data, then the bell curve would be this,



→ And this Right skewed Distribution is called Log Normal Distribution

→ Wealth Data.

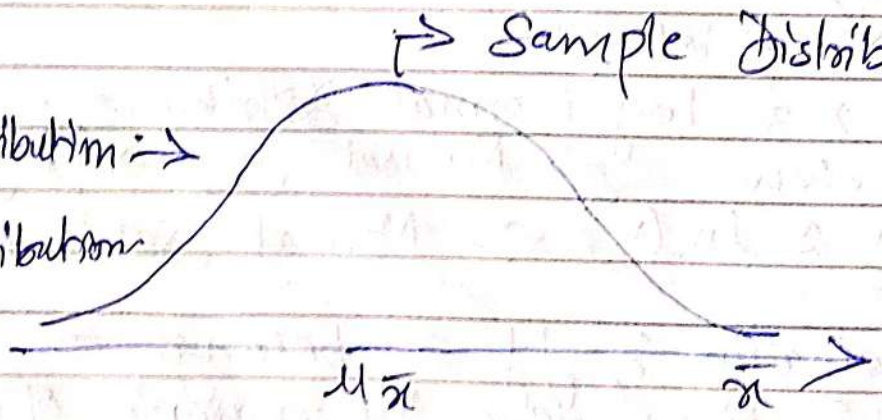
Sample size

↳ If our population is Normal / Gaussian Distribution or log-normal Distribution and if we apply some sample size (cond^n) randomly to the distribution data then will always get a Normal / Gaussian Distribution.

Sample size → Choosing random set of samples from the population data and using the set of sample Mean plotting the graph will always get a Normal / Gaussian Distribution.

So, Central Limit theorem tell that the Distribution is Gaussian or Normal Distribution or log Normal Distribution or whatever type of Distribution it may be, if we select a sample size which was $n \geq 30$ and we take Mean Sample Mean of each set of sample and plot it on the graph then it will always form Normal / Gaussian Distribution.

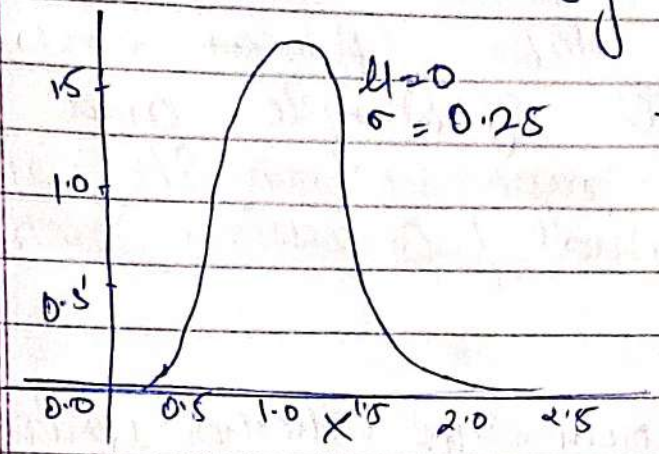
And this Central limit theorem says whether your distribution is Gaussian or not, it doesn't matter. ~~Because if it~~ because if it takes Sample Size condition of $n \geq 30$, then it will always give Normal / Gaussian Distribution.



Isse kehte Central limit theorem.

7.11 Log Normal Distribution.

will study → what is log Normal Distribution
→ What its properties
→ what its mathematical eqn
→ Examples of data which follow log Normal Distribution.



⇒ This is called Right Skewed and this is an important property of log normal Distribution.

It seems to be like Gaussian Distribution type, but the Right side is elongated.

let see this Distribution in Mathematical Way,

If $x \overset{\text{belong to}}{\sim}$ Log Normal Distribution.
then → Natural (log) → loge
$y \sim \ln(x) \Rightarrow$ Normal Distribution.

Explanation: If x belongs to log Normal Distribution. (Take the example of above distribution graph) and we consider that its a log Normal Distribution.

So, to verify it we take a new random variable (y) in which we have applied log to the base "e" ⇒ loge in every X's variable, so whatever output or if I plot it, I'll get Normal Distribution.

And if this property get satisfied, then we can say that X belongs to log Normal Distribution.

We can write that Algorithm inversely also :-

↳ $x \approx \exp(y)$ — $y \approx \ln(x)$ - Previous Algorithm

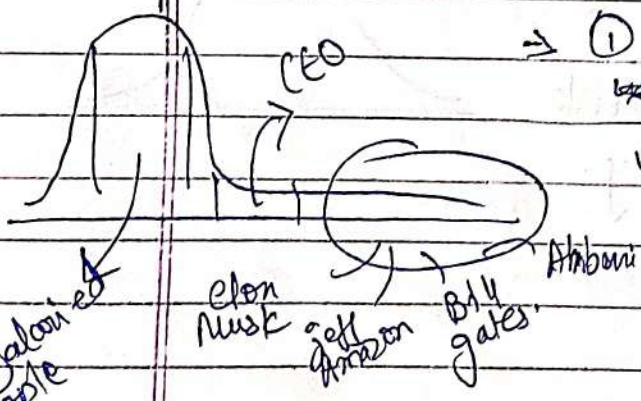
X value kojayega — These Both are Same

And if I inverse $\ln$ then Obviously it will become exponential of y $\exp(y)$ and X will become value.

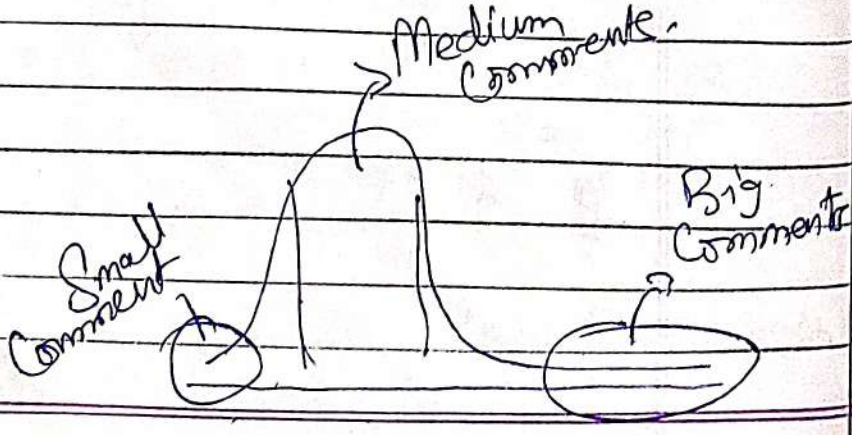
And these both will follow Normal Distribution if my X is log Normal Distribution.

Examples :- let see the examples.

⇒ ① Wealth Distribution
Very less people have more wealth.



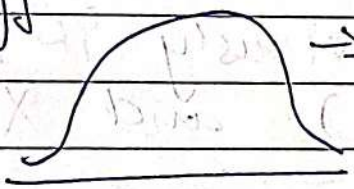
② Comments on Youtube

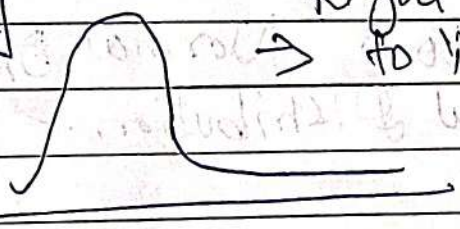


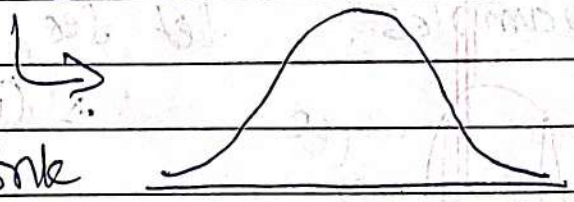
Now, Where we use this Log Normal Distr

(1) Machine learning Application
Whenever we in Machine learning Algorithm we use dataset

If I work with Simple Linear Regression

gj  → Normal Distribution, It will work efficiently.

gj  → Right skew, well apply log to its X's variable And it will convert it into a Normal Distribution

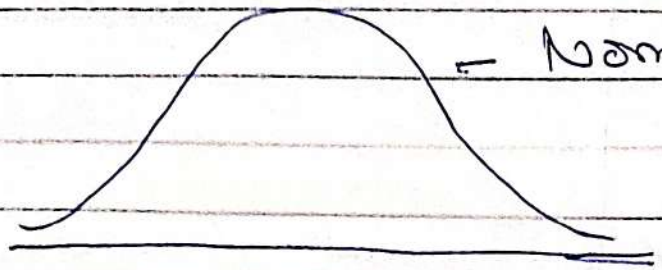


And then it will work efficiently or properly.

T.12 Power law Distribution.

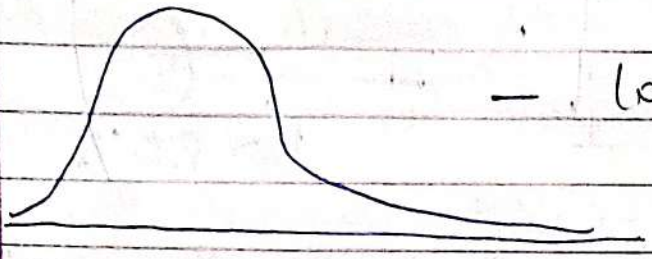
Page No.
Date

(1)



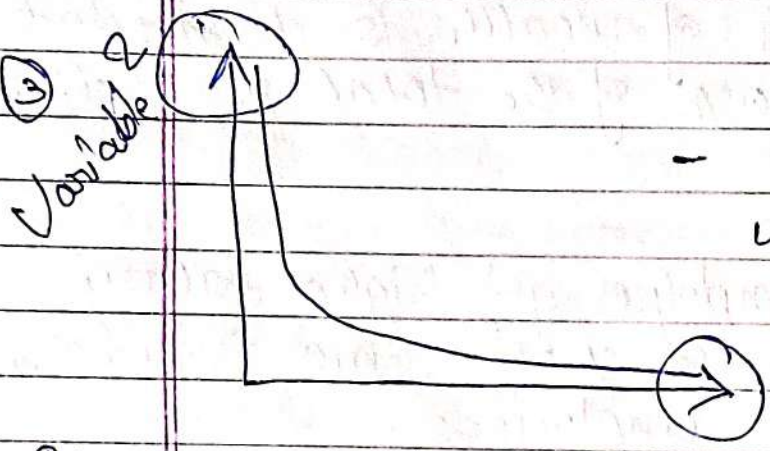
- Normal Distribution

(2)



- log Normal Distribution (Right Skewed)

(3)



- Power law Distribution, will see its examples later

If you want to know About Power law Distribution Quickly then always remember this Rule

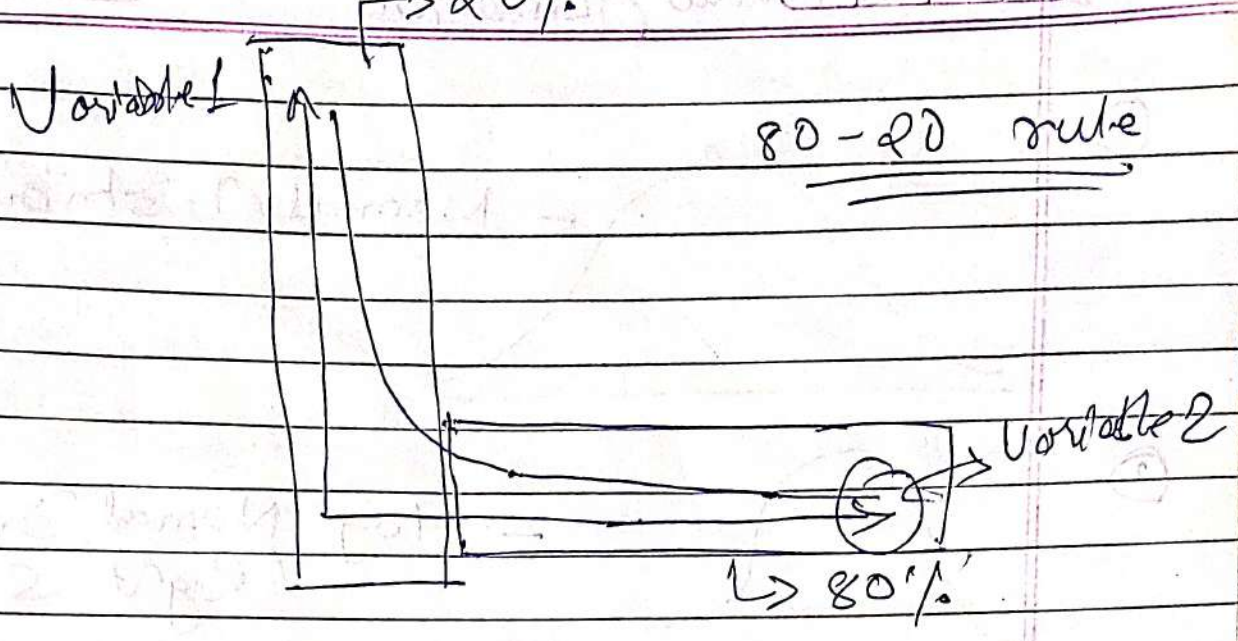
80 - 20% Rule

Both of these variable forms a relationship of 80-20% Rule.

Eg:- IPL → 20% of Team is responsible for winning 80% of match

explain:- If you take any match of IPL, jo 20% of Team honge, matlab 11 mai se 2 ya 3 log itna Acha khelte hai ki wohi match jeeta dete hai 80% of the match.

Now lets plot a power law Distribution



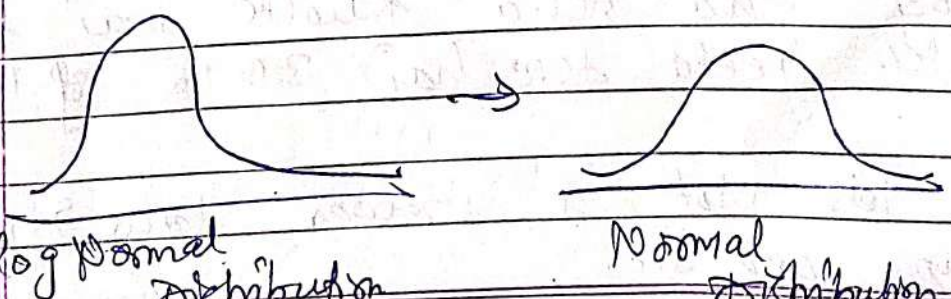
2nd example :- 80% of wealth is distributed with 20% of the total population

3rd Eg :- 20% of employees get higher salary with others than 80% of same Organisation employees.

4th Eg :- 80% of the expensive cars are bought by 20% of people in the world.

5th Eg :- Oil rich nation (Dubai, Saudi Arabia ... etc.)
↳ 80% of the total oil is with 20% of the nation.

Now, there is a Question that previously we used to log Normal histo → Normal Distribution By Applying $\log_e$ to the Variable (X)



So, Can we Convert this Power Law Distribution into Normal Distribution?

↳ Yes!

we Convert this power law Distribution into Normal Distribution Using Box Cox Transform. will see in Coming Videos.

Now, whoever follows this type of Distribution we call it as Pareto Distribution.

And this Pareto Distribution Using the concept of Power law Distribution distributes the Data.

will see Pareto Distribution Math on later Videos.

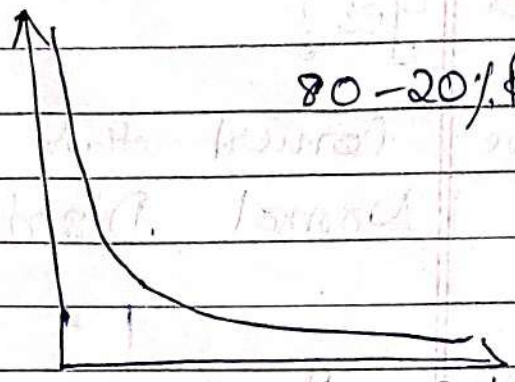
But let understand that Pareto Distribution is not a Normal or Gaussian Distribution, It is a non-gaussian Distribution which follows Power law Distribution

Pareto Distribution

Whenever when we discuss about Pareto Distribution, we should always power law Distribution

↳ To know about Power law Distribution

↳ Previous page.



Any Distribution which forms power law Distrib^n then we specifically called it as pareto Distribution

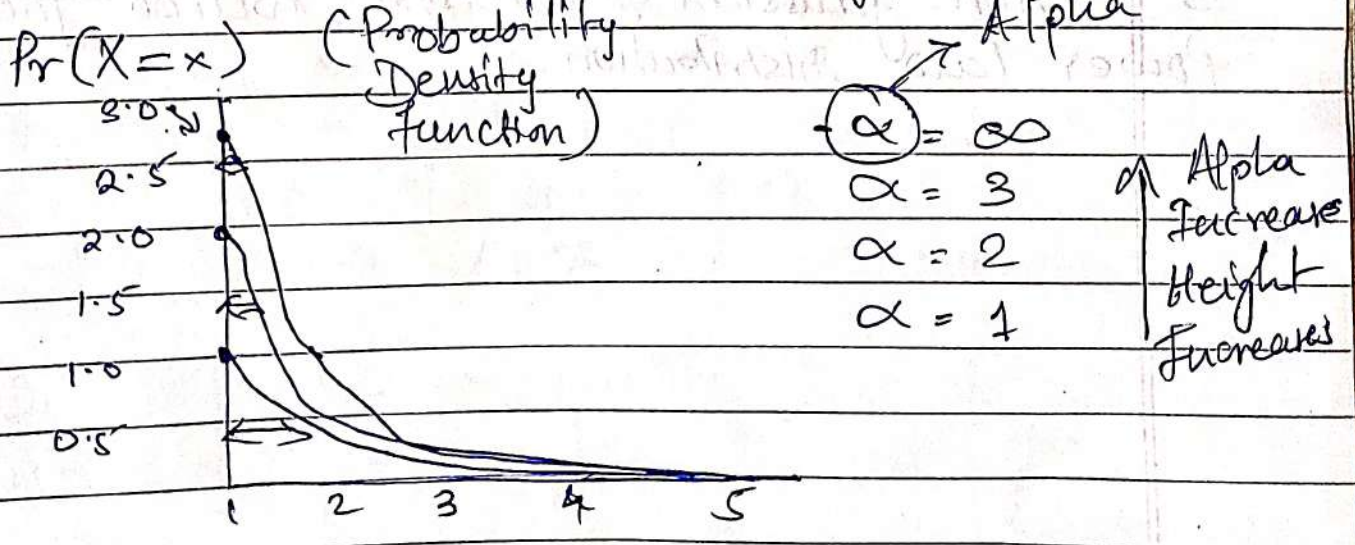
Interview Question

↳ What is Pareto Distribution?

Ans ⇒ It is an example of Power law Distribution and this Distribution are Non-Gaussian Distribution.

Let see Some Examples :-

Pareto Type 1



If Alpha Increases, the height also Increases. and if you notice the gap between the line and 1 gets reduced as the α Increases.

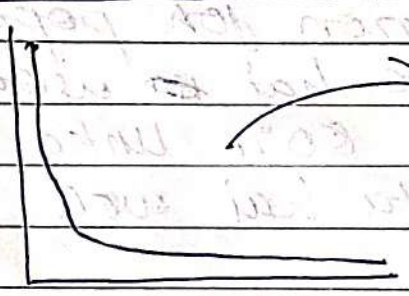
And It follows Power law Distribution rule.

Page No.
Date

Specifically, if I talk about the height and the gap, then it is determined using α (Alpha)

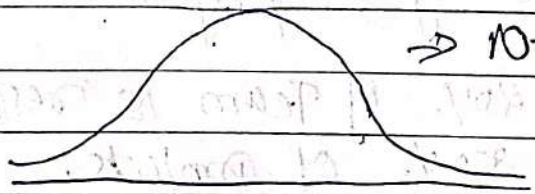
And why is it necessary to know, because if you train any model, like if I'm specifically working on Machine learning algorithms,

then some of the ML Algorithms do will say that my data should be in normal Distribution, ~~that~~ Lets assume if I have a dataset which follows this distribution which specifically follows Power law Distribution



Toh kya main isse ek normal Distribution main change kar paunga?

Can I change (Power law Distribution) into (normal Distribution)?

↳ 

⇒ Normal Distribution

Note ↓

The Importance of Normal Distribution is that Any of your Dataset forms Normal Distribution then it gets easy to work with Machine Algorithms like linear Regression, logistic Regression we can work on such models easily and can create efficient models.

So, if my Dataset follows the 80-20% Rule of Pareto Distribution and I want to convert that into Normal Distribution then the Answer for this is Yes!!
By using Box Cox Transformation
↳ we'll see this using Python Programming.

Eg:- of Pareto Distribution

↳ ① 20% of the product in Amazon is responsible for 80% of Sales.
↳ (Amazon ke popular products kitne hai ~~wo~~ woh 20% hote hai And 80% unka sales revenue hota hai woh yani ye ata hai)

② 20% of defects solves the 80% of upcoming defects.

③ 20% of Team is responsible for completing 80% of the Project.
↳ 20% of Team is responsible for delivering 80% of projects.

Hypothesis Testing

It's very important topics for the people those who are Approaching for the Interviews of Data Analyst, Data Science.

There is a huge chance of asking Question on Hypothesis Testing.

Defn: Hypothesis Testing is a form of Statistical Inference that uses data from a sample to draw conclusions about a population parameter or a population probability Distribution.

↳ wikipedia

Another Defn:- Hypothesis Testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

↳ Investopedia

And in Both these Definition, we are talking about Inferential Statistics

In, Inferential Statistics, will probably take Sample Data. And from this Sample Data whatever insights or Information we are getting, will come out with a conclusion about the Population Data.

To get the conclusion from the Sample Data will use Hypothesis Testing.

And In this Hypothesis Testing, there are different techniques :-

Z-Test
T-Test
Anova Test
Chi Square Test.

Let's Start with the steps of Hypothesis Testing.

↳ Let's do an experiment on the coin, if its fair or not

→ So I took a coin, where head & tail is its Outcome, So I want to find out if this coin is fair or not.

How fair?

① So, if we toss the coin and 50% of time I get head and another 50% of Time I get tail then we can say that the coin is fair.

② let say 60% of time - head
& 40% of time - tail
then also this coin is fair. we can call 60-40 ratio. it as fair

~~60%~~ ~~=~~ ~~50~~

③ 30% - head
70% - tail
then this coin is unfair

And All this Outcome we can find out through Hypothesis theorem.

Hypothesis testing Steps :-

Step ① → NULL Hypothesis (H_0) = Coin is fair.

② → Alternate Hypothesis (H_1) = Coin is not fair

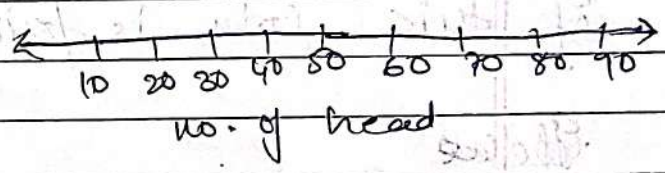
③ Experiment to prove ~~that~~ whether the NULL Hypothesis should be Accepted or rejected.

↳ Experiment : (Tossing coin for 100 times.) ← sample of Data

Coin is fair or not ← population.

① 1st 100 times
50 - head
50 - tail
Coin is fair

② 2nd 100 times
60 - head
40 - tail
SHU the coin is fair we can consider the 60 - 40 ratio.



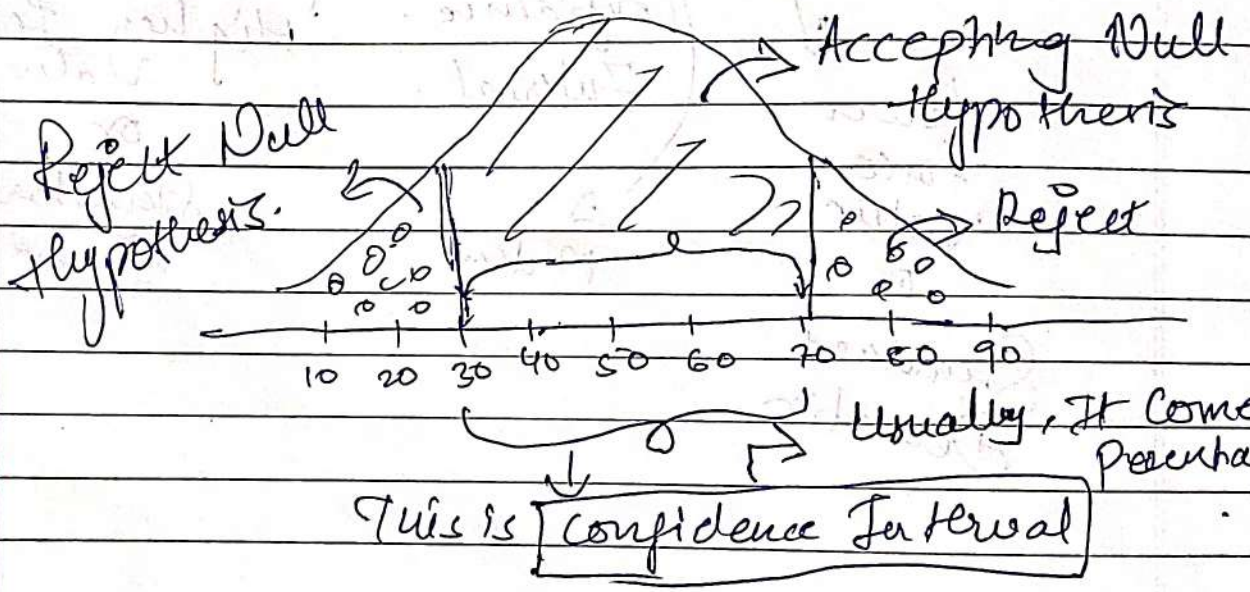
③ 3rd 100 times.
70 - head
30 - tail
Woo, Here I get conscious that I got 70 times head and remaining 30 times tail. So By this will probably the coin is not fair.

→ 2 Fair Test

Whenever, we toss the coin, there are possibilities of going below 50 or Above 50 ~~or else~~
& let say the Average is 50%.

let Assume, if we get 90 times head, so ~~to decide~~ we obviously confirm that the coin is not fair.

And for that we decide Confidence Interval



And how ~~to~~ Confidence Interval is defined by?
also

↳ Confidence Interval is defined by Domain Expert.

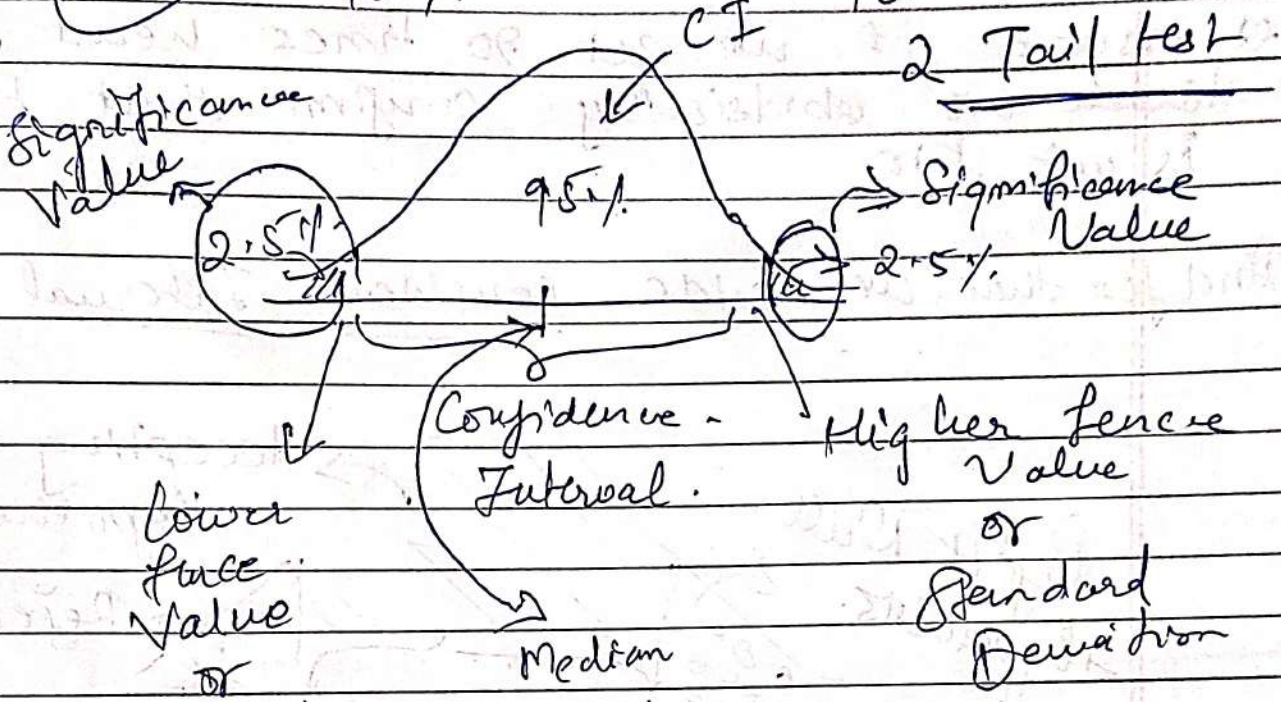
Domain Expert Eg: let Assume there is medical Usecase let say I want to take covid vaccination and I want to check whether this vaccination is Right or not, so will see its impact and effect ~~on~~ and according to that ~~will~~ ~~be~~ the Domain expert will decide its Confidence Interval.

Confidence Interval is also called as Decision

Note :-

Confiden Interval Usually comes in percentage.

C.I = 95%.



Lets take 2 Tail test.