Statistics :- it is an area of applied mathematics concern with the data collection, analysis, interpretation & presentation.

Measure of centre is statistical measurment and it's represents summery of Data. ie 'Measure of Central Tendency' → Distribution of Data ('Column')

We have three measures — 1) mean 2) median 3) mode.

Mean :- Measure of avg. of all the values in sample is called mean

Represented by μ ⇒ $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$ → No of observations.

Median :- Measure of Central value of the sample set is called median.

(1) ODD value → Median = $\left[\frac{n+1}{2}\right]^{th}$ value

Ex: [10, 20, 30, 40, 50] ⇒ '5' no of observations. = $\frac{5+1}{2} = \frac{6}{2} = 3^{rd}$

3rd position value is — '30' ⇒ Median.

(2) Even value → median = $\left[\frac{n}{2}\right]^{th} + \left[\frac{n}{2}+1\right]^{th} / 2 = \left[\frac{6}{2}\right] + \left[\frac{6}{2}+1\right] / 2$

[10, 20, 30, 40, 50, 60] = $(3)^{rd} + 4^{th} / 2 = \frac{30+40}{2} = 35$ → median

Diff b/w mean & median :- * Both are measure the central Tendency

mean → Gets Imputed with the precence of Outliers

median → Won't Get imputed with outliers.

mean → Time Complexity O(n) less for mean.

median → Time Complexity O(n log n) more for median.

	No outliers	Outliers	
~~Mean~~	— ✓	✗	} ~~Time Complexity~~.
~~median~~	— ✗	✓	

Mode :- is the value which is repeatedly occuring in a Dataset.

(ie) mostly occuring element in dataset

Ex : [1,2,3,4,4,3,4,2,5,7] → sort Step (1)

[1,2,2,3,3,4,4,4,5,7] — mode = 4 — Uni mode

[1,2,2,2,3,4,5,5,5,6,6,7] → mode = (3,5) — Bi mode

[1,2,3,4,5,5,6,6,7,8,8] → mode = (5,6,8) — Tri mode

Mean — when outliers impact — ✓ [Effected]

Median — when " " — × [Not Effected]

Mode — " outliers " — × [Not Effected]

Measure of Spread :- It describe how similar or varied data points for a particular variable.

We have (4) measures — ① Range — Max - Min.

② Inter Quartile Range.

③ Varience

④ Standard Deviation

Ex : [10,20,30,40,50] = [50-10] ⇒ 40. (Range)

↳ This Dataset is one column of Data So we called it Univariable. If we work on 2 columns it should Bivariable

So. Univariable we can perform (as well, Bi, multi also)

↳ 1) Measure of central tendency → 1) Mean (a)
2) median
3) mode

2) Measure of Spread.

↳ 1) Range ② IQR ③ Varience ④ Std. (6)

σ^2

Percentile :- it indicates the percentage of scores that fall below a particular value, They tell, where a score stands relative to other scores

Ex: If we attend an exam of 100 marks and we got 65 marks in it [65/100] and ranked as 1st then - 100% percentile if tells us no one or any value above the mark that i goted.

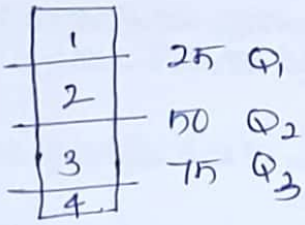
So here my score is 100th percentile.

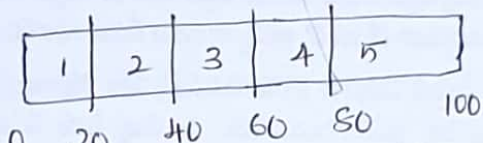
if any one got 75th percentile it means 75% of Data fall below that value or Rank and 25th percentile is above that value.

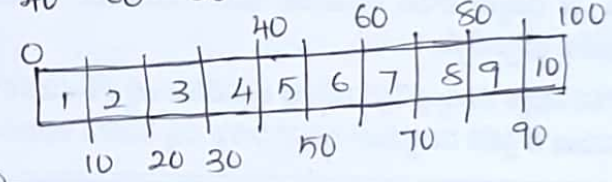
Quartile mean Data devided in to 4 equal no of parts.

Quentile mean " " 5 equal no of parts

Decile mean " " 10 equal no of parts







IQR = Median = $Q_3 - Q_1 \Rightarrow Q_2$

IQR :- Devide a rank in 4 equal parts Denoted as (Q_1, Q_2, Q_3)

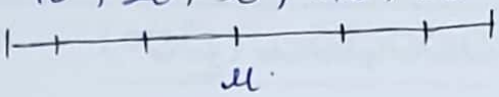
if we want IQR = $Q_3 - Q_1 = 75 - 25 = 50 \Rightarrow Q_2$

Simply we can say IQR is = 50th percentile.

* When outliers imputed on IQR that doesn't effect to IQR.

Varience :- is a measure of Spread., it tells how far away the Data point from mean in dataset.

Ex: 10, 20, 30, 40, 50 ⇒ mean = 30. = μ.



30-10 = 20.
30-20 = 10.
30-30 = 0
30-40 = -10
30-50 = -20
} points.

Varience represents in (σ^2).

$$\sigma^2 = \frac{\sum_{i=1}^{n} |\mu - Obj_i|}{n}$$

Standard Deviation : is a Square root of ~~Standard~~ Varience

Denoted as (σ) = $\sqrt{\text{Varience}}$.

M. of Spread
- Range – when outliers imputed – ✓ (Effected)
- IQR – when outlier " – × (Not Effected)
- varience – " outliers " – ✓ (Effect)
- Std. – " outliers – ✓ (Effect)
- MAD – mean absolute Deviation – " – × (Not E)

Mean A. Deviation :-

it is the avg distance b/w each data value and mean.

Median Absolute Deviation :- it shows the how far away the data points from median in Data set.

$$M.A.D = \text{Median} \mid \mu - Obs_i \mid$$

In this it rejects all outliers

* Standard Deviation gives Distribution of variables.

By above all topics we covered on Uni variables which perform

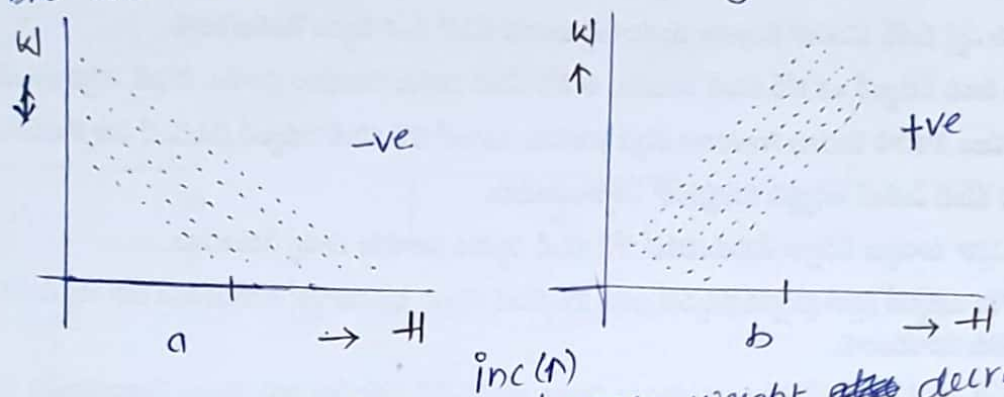
Uni variable
- Measure of Spread
- Measure of Central Tendency

M. Relation :-

On Bi variables we find → Co-variance, Co-Relation } mesure of relation.

* Univariables ie deals only one column, or one variable where Bivariable means deal on two variables or two columns

if we have a data set on which height & weight present- if we plot a graph b/w this two variables (H, W), if the graph is like this -



inc (↑)

In fig a when height ~~increase~~ ~~decrease~~ weight ~~also~~ decrease (↓) this tells the Correlation b/w two variables is -ve Correlation ie Negative Cov

if fig b observe height when increase weight also increases this tells relation b/w two variables is positive ie Covariance is positive.

if we Observe variance of Univariable $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - x_i)^2}{n}$

Similarly in Bi-variables $Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - x_i)(y_i - y_i)}{n}$

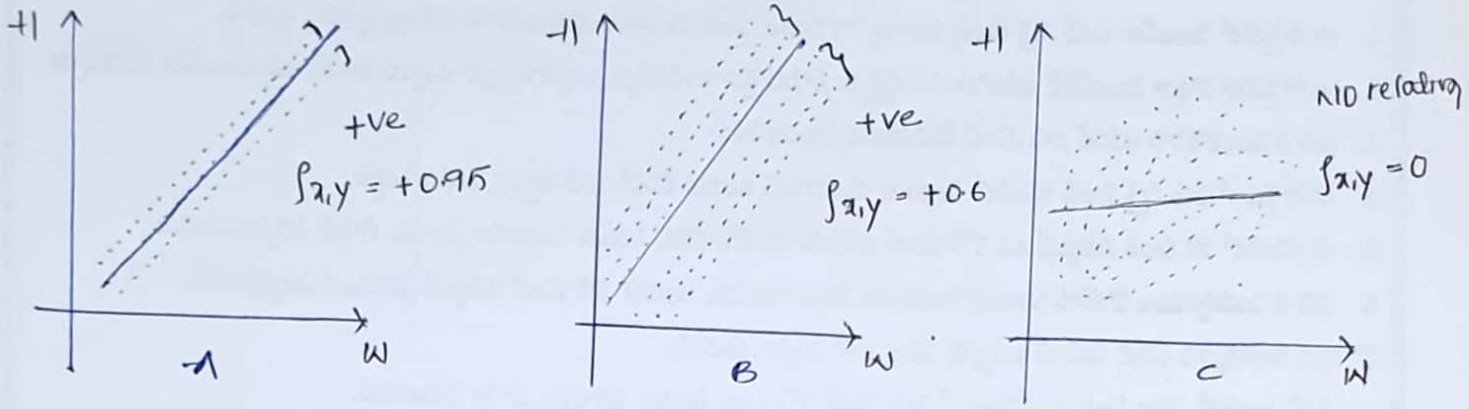
which tells the sum of diff b/w mean-w each data point in a axis and mean on 'y' variable-w each point.

Correlation :- it measures the degree to which, two variables move in relation to each other in other it call as pearson Correlation Coefficent and denoted as $\rho_{x,y}$ ∴

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

The correlation value always lies in b/w $\boxed{-1 \leq \rho_{x,y} \leq +1}$

it also similar as covariance when we plot figures as follows-



Above figures tells us there is positive corelation b/w two variables height & weight $\rho_{x,y}$ is +ve ie when Height increase weight variable also inc., when height dec, weight also dec thats -ve corelation.

fig (c) tells there NO relation. ie correlation is zero.

Diff b/w covariance and correlation :-

Coverience is nothing but measurement of correlation, it indicates the directions of the linear relationship b/w variables

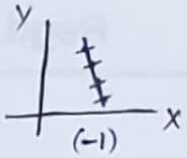
Correlation on the other hand measures both the strength and direction of the linear relationship b/w two variables

Correlation coefficient support linear relationship.

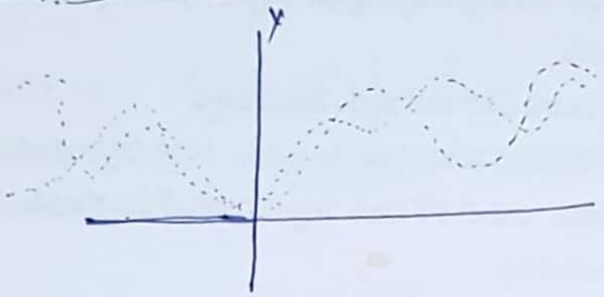
if we observe figures A, B. both tells (+ve) co-relation but Strength of the relation is changed, cov doesn't tells Strength

Both tells us direction, linear relationship of two variables.

Correlation of (-1) indicates a perfect negative correlation that as one variable inc the otherone is decreases in this case all the data points lies on one line



Problem with Pearson Correlation Coefficient :-



→ Sine Sodal data. varies like this

P. correlation coefficient only tell Linear Relationship won't tell Non-linear R.

In. L.R only, $\rho_{x,y}$ = values in other wise it fails ie $\rho_{x,y} = 0$.

Probability :- it is Study of Uncertainity.

Random Experiment :- it is a process for which Outcomes can not be predicted with Certainity.

Ex: Tossing a coin, Rolling a dice, picking an object

Sample Space :- Set of all the possiblity of Random Experiment

Ex: Tossing a coin — {H, T} ⇒ S.S.

Rollig Dice — {1, 2, 3, 4, 5, 6}

Event :- Any Subset of Sample Space

Ex: Tossing Two coins. ⇒ Random Experiment

SS ⇒ {HH, HT, TT, TH}

Event ⇒ E_1 — {HT, TH} - Getting Exactly 1 Head ⇒ 1/2

E_2 — {HH, TT} - Something on Both Toss ⇒ 1/2

E_3 – {HH} – Getting Two heads ⇒ 1/4

E_4 – {TT} – " Two Tails ⇒ 1/4

$$\text{Probability} = \frac{\text{Favorable Outcomes}}{\text{Total Outcomes}}$$

In case of $E_1 \Rightarrow P(E_1) = \frac{\text{Size of Event}}{\text{Size of Samplespace}} = \frac{2}{4} = \frac{1}{2}$,

Rules/Axioms of probability :-

① Probability of an Event always lies b/w $0 \le P(E) \le 1$

because possible Events is Subset of Sample space ie Event value is Under the Sample Space. not beyond S.S.

② Probability of Sample Space is always 1 $P(S.S) = 1$

③ for any Sequence of 'Events' that are mutually Exclusive

Ex: $S_1 = \{1,2,3\}$, $S_2 = \{3,4,5\}$, $S_3 = \{4,5,6\}$

S_1, S_2 = (1, 2, 3) (4, 5) — common value = 3 — Not mutually Exclusive

S_1, S_3 = (1, 2, 3) (4, 5, 6) → M. Exclusive

By the above we can Say —

$$E_1 \cap E_2 \cap E_3 \cdots \cap E_n = \phi \rightarrow \text{M. Exclusive}$$

$$P(E_1 \cup E_2 \cup E_3 \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$$

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} E_i$$

Ex: When Rolling Two coins — E_1 – Both Head – {HH} ⇒ $P(E_1) = \frac{1}{4}$

Events like — E_2 – {TT} = 1/4

E_3 – {HH, TT} – $P(E_3) = 1/2$,

if we observe Event 3 $P(E_3)$ that belong Union E_1 & E_2

simply — $E_3 = E_1 \cup E_2 \Rightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

$P(E_3) = \boxed{P(E_1 \cup E_2) = P(E_1) + (P(E_2)}$ if $\rightarrow P(E_1 \cap E_2) = \phi$ ↓ represents '0'

$P(E_3) = P(\{HH\}) + P(\{TT\}) \Rightarrow \frac{1}{4} + \frac{1}{4}$

$P(E_3) = \frac{2}{4} \Rightarrow 1/2$.

if $P(E_1 \cap E_2) \neq 0 \Rightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

Ex: What is the probability of getting prime no's when rolling a die?

→ Step 1 — R.E — Rolling Die (Getting prime no's is Event)

S.S — $\{1,2,3,4,5,6\}$

Event — Getting prime no $\{2,3,5\}$

$P(E) = \frac{3}{6} \Rightarrow 1/2$

if we want probability of getting even no's in rolling die.

Event = Getting Even $\{2,4,6\}$

Prime no's $\{2,3,5\}$

In this Condition's Cases we use Conditional probability.

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \Rightarrow = 1/3$$

if we take Event getting Even no's as → A $P(A) = 3$

Same $P(B) = 3 \Rightarrow P(A \cap B) = (\{2,3,5\} \cap \{2,4,6\})$

$P(A \cap B) = \{2\}$.

All possible Out Comes = 1 only

Sample Space = 6 $\Rightarrow P(A \cap B) = 1/6$

Simply $\Rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3}$

(01) – $P(A|B)$ – Conditional prob – P(Getting prime | Even no)

$P(B) = P(\{2,4,6\}) = \frac{3}{6} = 1/2$

$P(A|B) = \frac{1/6}{1/2} \Rightarrow 2/3 = \frac{1}{3}$.

CRISP-DM FRAME WORK :-

Cross Industry Standard process for data mining.

① Business Understanding → it means where the data comes from if there is no data to perform EDA, then you have to go web scrapping or go to database and try to fetch the data. Once data comes then we have to understand which type of data, business problems we have to solve.

② Data Understanding by EDA, Analysis, Numpy, statistics, Vizualization.

③ Data Preparation → Treat Outliers, missing values, encoding data, convert Data.

④ Machine Learning modeling.

⑤ Evaluation of M.L Model

⑥ Deployement

Random variable :- "X" it is a variable whose possible values are numerical Outcomes of random phenomenon.

Ex: Tossing a Coin.

Random Experiment – Tossing Coin

Sample Space – {H, T}

Random variable – 'X' – Counting the no of heads – {H, T} – {1, 0}

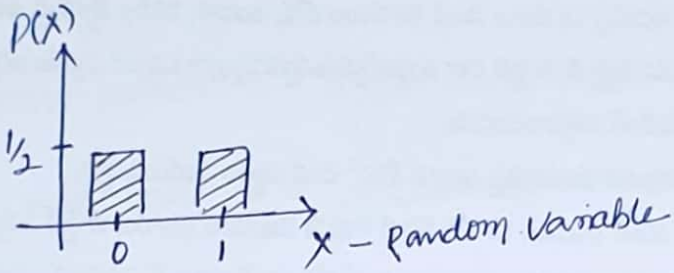
Computation – Compute the probability of 'x' ≃ p.d.f

$P(X=0) = P(\{T\}) = 1/2$

$P(X=1) = P(\{H\}) = 1/2$

$\begin{bmatrix} H \\ T \end{bmatrix} \begin{matrix} \to 1 \\ \to 0 \end{matrix}$

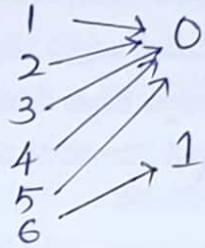
Ploting – Plot the probability distribution function (p.d.f)



Ex-2 R.E – Rolling a dice

S.S – {1, 2, 3, 4, 5, 6}

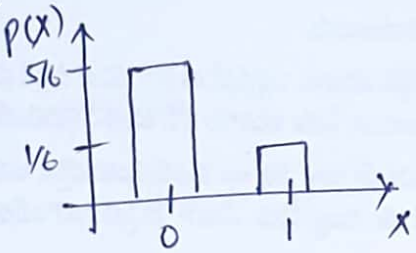
R.V – Getting a '6' : X



p.d – $P(X) \Rightarrow P(X=0) \Rightarrow \{1,2,3,4,5\} = 5/6$

$P(X=1) \Rightarrow \{6\} = 1/6$

Plotting –



* There is more chance of getting failure than success

Simply we say Random variable is a function which will take Sample Space as input & give some o/p in a form of a number

$X : \{H, T\} \to \{1, 0\} \Rightarrow$ o/p

↳ S.S.

Classification :–

- Random variables
  - Discreate Random variable
    - Bernolli
    - Binomial
  - Continious Random variable
    - Uniform
    - Normal

displot → distribution plot it consist of three things.

① K.de → p.d.f
② Histogram.
③ c.d.f

This function distplot is discontinued in the future we have to use a new function called as. "displot"

Discreate Random variable :- it has a countable no of possible values. the Probability of each value of a discreate random variable is between 0 and 1 and Sum of the probabilities equal to 1.

Bernolli Random variable :- if the outcome taken Countable values (finite) set aswell Random variables containing two values only (ie Binary set of values) is called Bernolli Random variable.

Ex: {0,1} => Here '0' represents failure and '1' is Successor. this variable Contains only two values that ase finale number and having only two values 1) failure, 2) Successor. Such variable is called as "B.R.V"

P.M.F → Probability Mass-function is used when we have descreate Random variables it tells as variable ie Bernolli or not

P.M.F ⇒ P(X=1) ⇒ P — Success
P(X=0) ⇒ 1-P — Failure

Binomial Random variable :-

it is a collection of bernoli Random Variables (X) is called as Binomial. R.V.

Ex: if we Tossing a coin it Containe $\{H, T\}$

S/F S/F

the Combination of Success/failure is Simply Binomial R.V.

1. R.E — Tossing Two Coins.
2. S.S — $\{HH, HT, TH, TT\}$
3. R.V — Counting the no of heads.

HH — 2
HT — 1
TH — 1
TT — 0

$\{HH, HT, TT\} = \{2, 1, 0\}$ → Shows Density of a variables.

4. Compute the probability of R.V. ⇒

$P(X=0) = P(\{T,T\}) = 1/4$

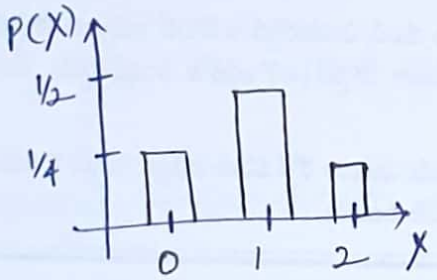
$P(X=1) = P(\{HT, TH\}) = 2/4$

$P(X=2) = P(\{HH\}) = 1/4$

5. plot PDF

⇒ $P(\{0,1,2\})$ → is not a Bernolli but it is a collection of bernolli.

⇒ Simply called as Binomial.



P.M.F of Binomial Random V. :-

$$P(X=i) = (p)^i * (1-p)^{n-i} \Rightarrow {}^nC_i = \frac{n!}{i!(n-i)!}$$

i = no of Success

p = probability of Success

$1-p$ = probability of failure

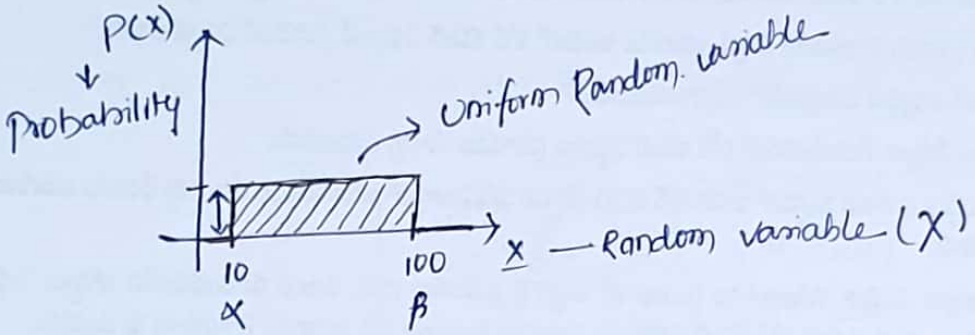
n = no of trials.

Continious Random Variable :- Uniform, Normal

'X' is a C.R.V variable, if the set of possible values are infinet or Uncountable is called as Continious Random Variable

Uniform Random variables :-

The probability that Uniformly distributed random variable falls with any interval of fixed length



$$P(X = x) = \frac{1}{\beta - \alpha}, \text{ if } \alpha \leq x \leq \beta$$

$$0, \text{ Otherwise}$$

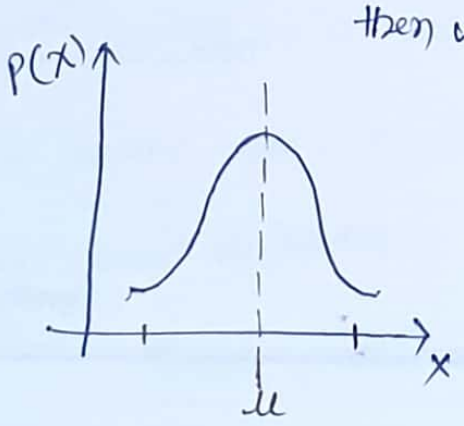
① R.V x is lies inbetween (α, β) where α is min value, β is max

Then the prob of R.V $(P(X) = \frac{1}{\beta - \alpha}$ otherwise $\rightarrow$ '0'

Normal Random variable :- if we plot a distribution plot (displot (kde=T

The plot looks like _ in bell curve & mirror image each other

then we called as "N.R.V"



$$R.V = X \sim N(\mu, \sigma^2)$$

N.R.V follows mean (μ) & Std σ^2

$$P.d.f \Rightarrow P(x) = \frac{1}{\sqrt{2\pi}\sigma} * e^{-(x-\mu)^2/2\sigma^2}$$

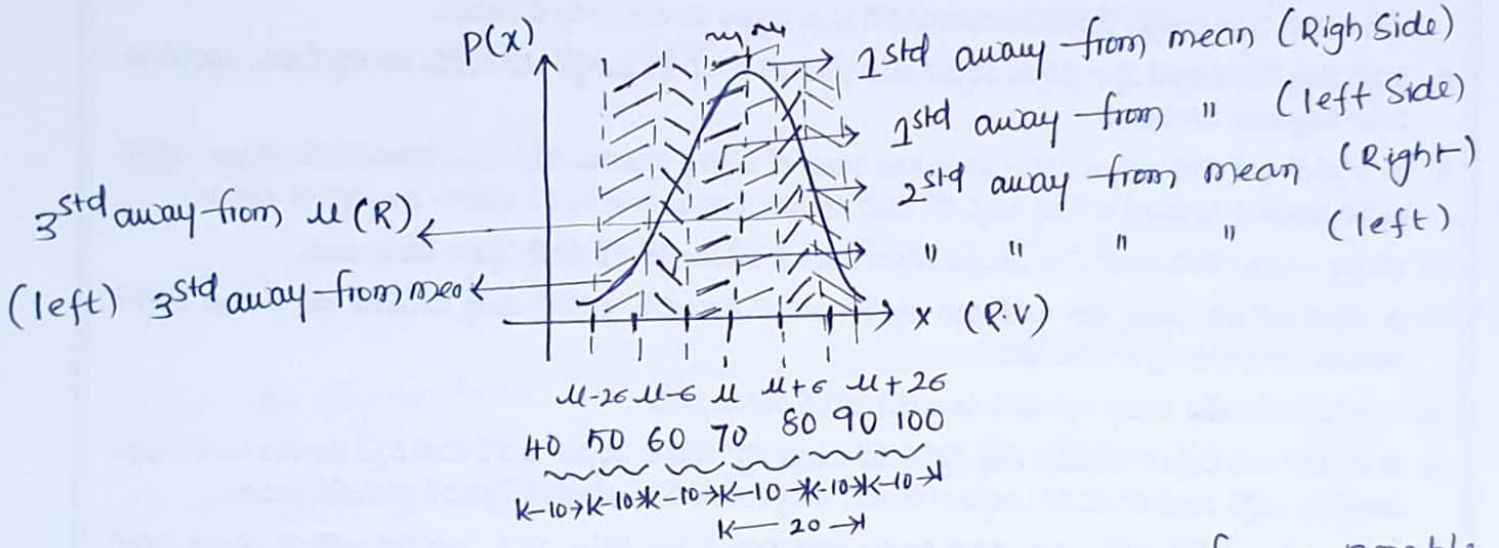
Normal Distribution :- This distribution also called as "Gaussion D."

it is a probability distribution, that is Symmetric about mean, Showing the data near the mean, in graph-form this distribution will appear as bell Curve., it follows mean, std.

There is imp role → 68-95-99.7%

Ex : if data distributed like normal and mean is 70 and std is 10

$$X \sim (70, 100)N \Rightarrow \mu = 70 \,,\ \sigma^2 = 100 \,,\ \sigma = 10$$



Simply we can say that 1 std lies 68% of data from a variable ie [60,80] lies 68% of data.

2nd std. lies 95% of data from variable ie [50,90] ~ 95%

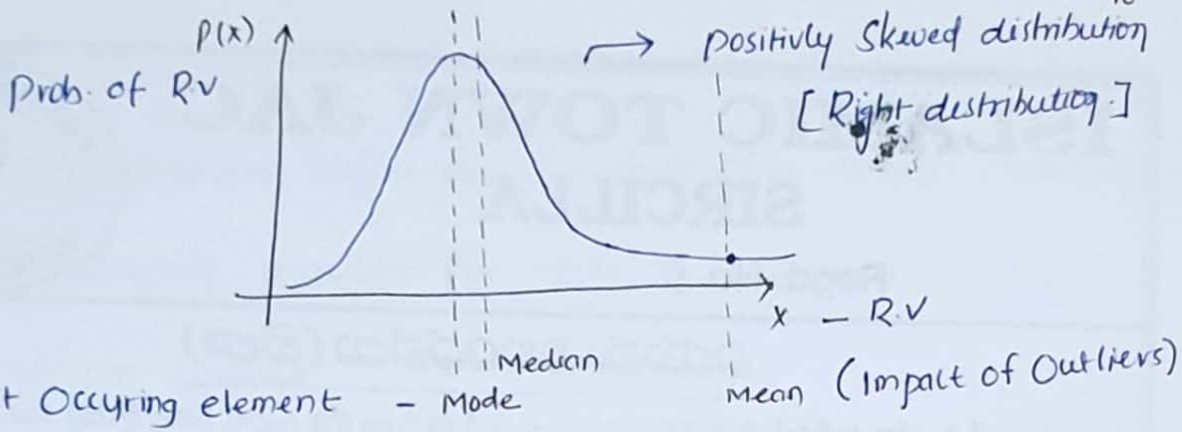
3 std lies 99.7% of data from variable ie [40,100] ~ 99.7%

1.Q: what happend to mean median mode, when you have positively or Negtively Skewed distribution ?

① **Mode ≤ median < mean**

μ is very largest as Compare to other, Mode, median. very near

2.



When we observe Salary, houseprices, Carprices in that we get this type of distribution, according to my observation mean (μ) get impacted with the presence of Outlier and median, mode are very near, in some cases same ie mean is very largest and median, mode are very close to each other.

Skewness :- it refers to distribution or a symmetry that deviates from the Symmetrical bell Curve, or normal distribution, in a set of data. if the Curve is shifted to the left or to the right, it is said to be Skewed. Skewness Can be Quantified as a representation of the extent to which a given distribution varies from a normal distribution. a normal distribution has a skew of zero.

Skewness value lies in between -0.5 to +0.5 we called as "Symmetric" if the Skewness of a variable is lessthan -0.5 or more than +0.5 it is not Symmetric.

if skewness value lies in b/w -1 to +1 than it is called as heavily negitively skewed and heavily positively skewed.

kurtosis :- it is a statistical measure. that defines how heavily the tails of distribution differ from the tails of a normal distribution kurtosis identify weather the tails of a given distribution contain extreme values.

for normal distribution kurtosis value is '0', if kurtosis value positive its looklike peaked, if the value is negative its look like flat.

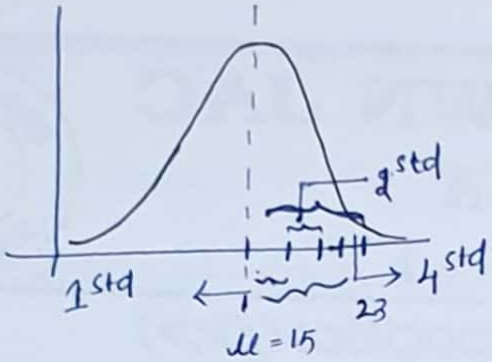
Standard Normal Distribution :- A Random variable (if it is) follow normal distribution with the mean of zero and standard deviation is one then this type of variable is called as standard normal variable, and the distribution known as standard Normal distribution.

$$X \sim N(\mu = 0, \sigma^2 = 1)$$

Random variable ↳ Normal Distribution

Z-score :- it is a measure of howmany std away ~~from~~ is a point from the mean. called as Z-score

Ex: if a variable having mean (μ) is 15 and std is 2, if we draw the distribution plot of that variable, if we take any value of variable in case '23', then this point how many std away from mean $\mu = 15$ is 4 std away, this 4 represent Z-score of the variable



$\Rightarrow \mu = 15, \sigma^2 = 4$

$\sigma = 2$

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

Q_i – point in variable

$$Z\text{-score} = \frac{23-15}{2} = 4$$

Distributions & Transformations :-

Statistics
- Univariet → M of central Tendency – Mean / Med / Mod
  - → M of Spread – Range / IQR
- Bivariet → M of Relationships – Cov, Correlation

Probability → Basics – R.E, S.S, Event
- → Axioms of probability.
- → Conditional probability → R.E.
- → Random variable → S.S / R.V
  - → p.d – Descrete → PMF. → Prob. distribution
  - → pd – Continious → PdF → ploting p.d.

if a given variable is normally distributed or not, ie (normal or not) to check this we used a test called as Q-Q plot.

Quantile – Quantile (Q-Q) plot :- (test of Normality)

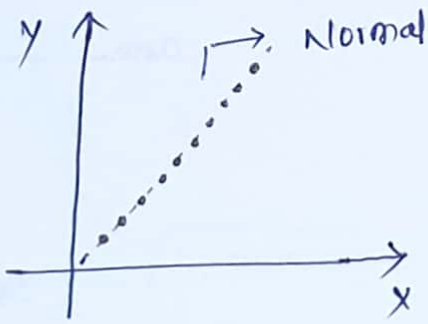


fig (1)

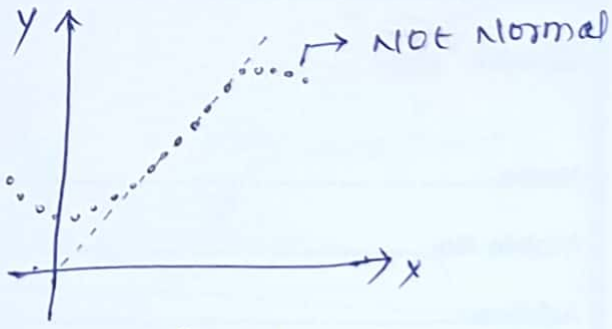


fig (2)

if we plot a Graph b/w two diffrent variables, all the data points lying on a same line (line = 45° in graph) then we called as the distribution is normal distribution. in case there is any distrabances like data points distracted away from line then we say they are not distributed as normal.

How to get a QQ plot :-

① we have to check ① Theoritical Quantity — Normal distribution. ($\mu = 0, \sigma = 1$)

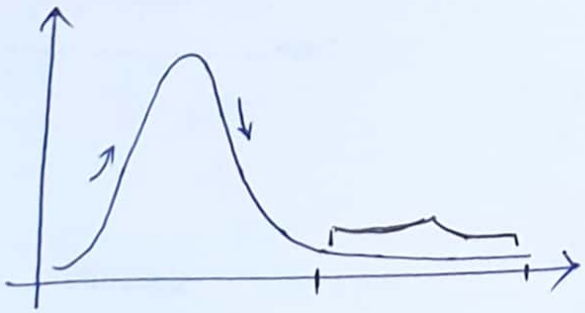
② Observed Quantity — Given distribution / column.

② Sort the values of Th.Q. and Obs.Q.

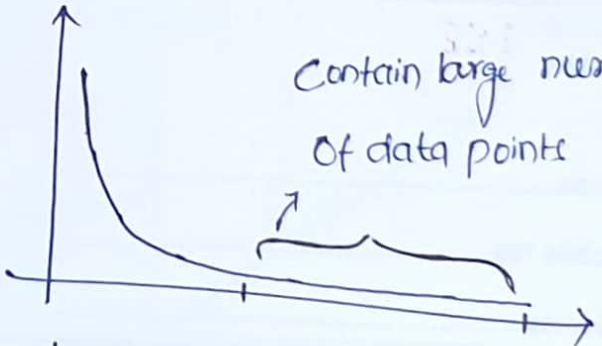
③ After Sorting plot a Scatterplot and take Observed Quantity (Column) if all the data points lying in a 45° line then we say Normally distributed. Otherwise not distributed normally.

Theoretical Quantity always lying 'N.Dist' because we use this "np.random.normal (loc = 0, scale = 1)"

Log Normal Distribution :-



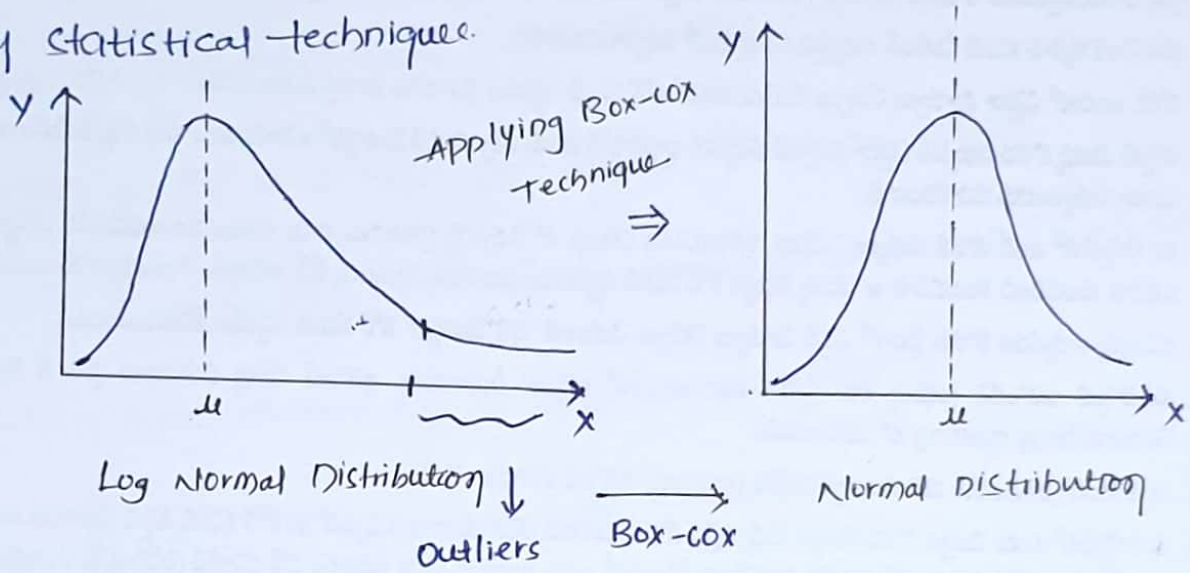
Pareto Distribution :-



if we draw graph on Salary, houseprice, carprice etc columns we get this type of distribution.

If we observe this two distributions it contains more Outliers the magnitude of outlier is high so it is impossible to delete the column because lot of data gonna [ie predict wrong analysis]

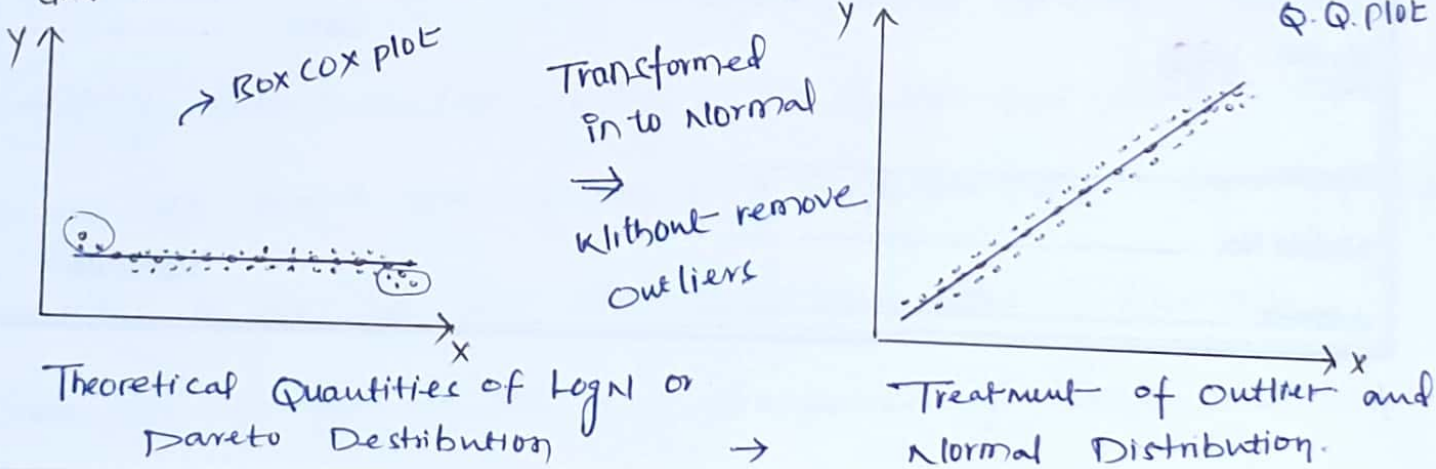
Box-Cox Transformation :- it is a transformation technique of a non-normal dependent variables like (Log Normal D, pareto D) in to normal shape, Normality is an important assumption for many statistical techniques.



→ Some more Distributions — Log Normal, pareto

→ Q.Q plot → using for test of Normality.

→ Box Cox → Transformation (Convert log.N, pareto -to Normal Distribution and treat the Outliers)



Inferential Statistics :- Estimating or Guessing.

it is dealing with forming Inferences & predictions about population based on sample of data taken from population. by using point estimation, we can determind the Inferential Statistics.

Point Estimation :- if we want to find the mean of (μ) population, firstly we draw a sample and then find sample mean ($\bar{x}$). and its estimate the population mean ($\mu \approx \bar{x}$) this is called point estimation

Population :- it is a set of similar items or events

Sample :- it is a set of uniformly, randomly selected items or events

population mean denoted by 'μ' and sample mean $\bar{x}$

$$\mu = \frac{\sum_{i=1}^{N} Obs_i}{N} \quad ; \quad \text{Population Variance } \sigma^2 = \frac{\sum_{i=1}^{N} (Obs_i - \mu)^2}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} Obs_i}{n} \quad ; \quad \text{Sample variance. } s^2 = \frac{\sum_{i=1}^{n} (Obs_i - \bar{x})^2}{n-1}$$

Note: (n-1) cause of bias in U.R.S (∵ Remove Impact of bias)

We have diff types of Sampling Techniques Such as Convienent, Valuentree Sampling, Uniform random Sampling, in this Convient and valuentree can't be give you perfect value of mean (ie $\bar{x}$ Sample mean $\mu_{\bar{x}}$ = Population μ_{pop})

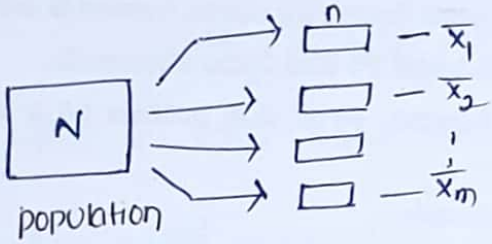
In U.R.S is gives you some what approx mean value. it have also some bias for that we put (-1) in denomenator of Sample variance

Centrel Limit Theorem :- if we have large papulation and we decided into many samples then the mean of all the samples from the papulation will be almost equal to mean of the entire papulation ie each of the Sample normally distributed

$\mu_{\bar{x}} \simeq \mu_{pop}$ → Mean of papulation.

↳ mean of sample

if we have population N and we are selected samples as Uniformly, Randomly distribution. if we take Sampling distribution ($\bar{x}$) it follows Normal distribution $\bar{x} \sim N(\mu_{\bar{x}}, \frac{\sigma^2}{n})$ according to Central limit theorem



Sampling Distribution $\bar{x}$, it follows Normal Distribution.

C.L.T :— 1. Sample Distribution mean $\mu_{\bar{x}}$ = population Mean μ_{pop}

2) Sample D. varience = $\frac{\sigma^2}{n}$ { n - size of each Sample. σ^2 - Pop varience.

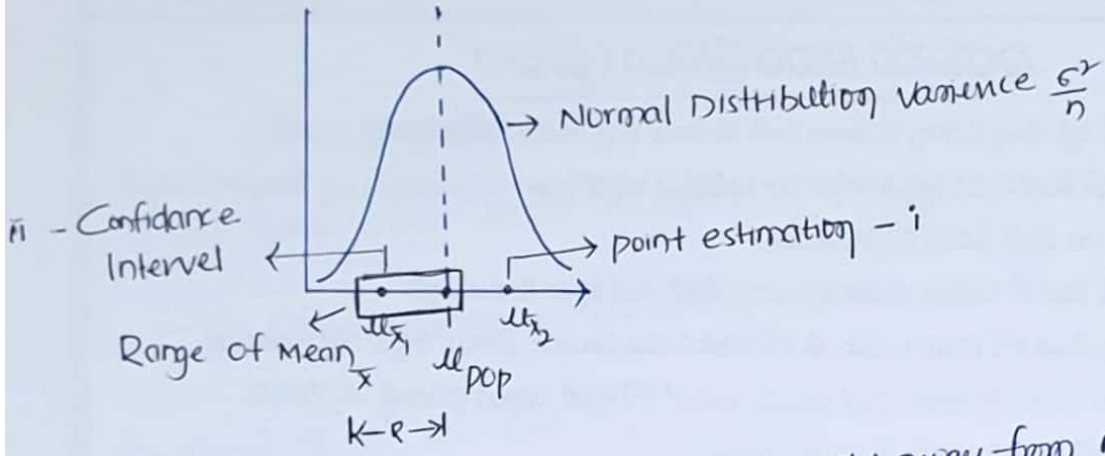
3) Sample D. Std = $\frac{\sigma}{\sqrt{n}}$ ⇒ also called as Standard Error

4) Sample . D. $\sim N(\mu_{\bar{x}}, \sigma^2/n)$.

After Sampling distribution every Samples mean [$\mu_{\bar{x}} = \mu_{pop}$] mean of population So if we consider one Sample from Sample distribution the Sample can be biased or convinent / valentary Sample and point estimates thats why we have to go Confidence Interval.

1) $\mu \approx \bar{x}$ → point Estimate, Sample can contain bias

2) C interval → $\mu \approx [\bar{x}_1 \pm$ Something$]$ with y%. Confidance.



⇒ interval may represent how many std away from mean to the point.

So that represent in z^* - score value then —

$\mu \approx [\bar{x}_1 \pm z^* \frac{\sigma}{\sqrt{n}}]$ with y% of Confidence. (std of pop is given)

if we find z-score value then we easily represent how much % Confidance

if $z^* \rightarrow 1^{std}$ away from mean it represent - 68% of Confidance

Ex: Consider one college, in that Students got jobs and their Salary package ~~listed~~ of some of the Students listed as follows [3.5, 6, 2.5, 3.5, 3.7] LPA

if we Calculate Mean of it $\mu_{\bar{x}} = 3.84$.

In this case we don't know the Salary package of every Student, who got jobs So what ever Salaries we are taken are Convienence Samples.

So $\mu_{\bar{x}} = 3.84$ is point estimate, Biased. So we need to go Confidance interval in this $[\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}]$ with n% Confidance., we don't know the value of "σ_{pop}", ~~dont know~~ μ_{pop} because we take Sample of values

Such case — we have to calculate S^2 = Sample variance

So in this case we have mean of sample so — easily we can calculate

$$S^2 = \frac{(3.84-3.5)^2+(3.84-6)^2+ \dots +(3.84-3.7)^2}{5-1} \quad n=5$$

$$S^2 = 1.1586$$

if we have don't know the value of σ_{pop} instead of it we can use

$S \rightarrow$ Sample standard deviation; if we use it. z^* value also changed.

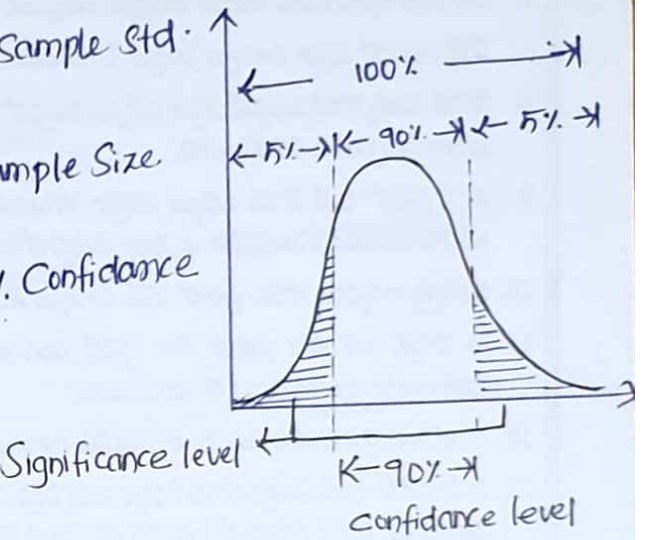
Instead of z^* we use $t_{n-1,\alpha/2}$ and it's called as T-score.

$$= \left[\bar{x} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Sample std.

↳ S.mean ↳ Sample Size

$$\left\{3.84 \pm t_{4,\alpha/2}\frac{1.1586}{\sqrt{5}}\right\} \text{ with } 90\% \text{ Confidance}$$



So Here 90% can be represent as 0.9

$\Rightarrow 1-0.9 = 0.1 \simeq 10\%$

Here 10% → Significance Level (α)

In $t_{n-1,\alpha/2}$ → Critical value / α - Significance level.

↳ Degrees of freedom ∴ Critical.V = $\frac{\text{Sognificance}}{2}$

D.O.F = Samplesize −1. ; C.V = $\frac{10\%}{2} = 5\% \simeq 0.05$

So — $t_{4,0.05}$ → we have to find the value from T-score

In T score represent D.OF., C.V $\Rightarrow T_{4,0.05} = 2.132$

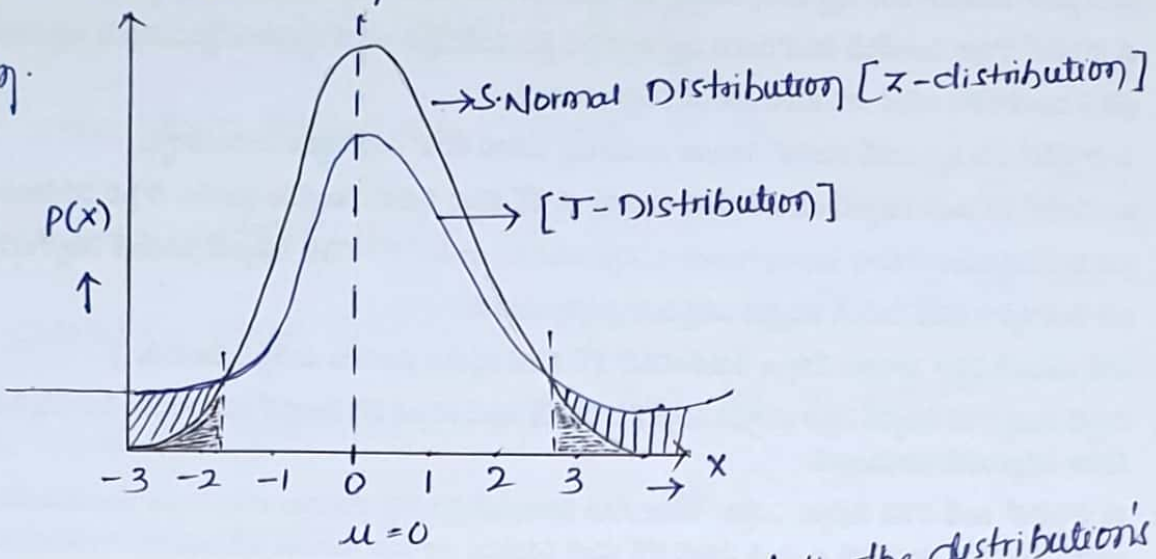
put all values — $\left[3.84 \pm 2.132 \times \frac{1.1586}{\sqrt{5}}\right]$ with 90% confid-

$[3.84 \pm 4.7] \Rightarrow [-0.86, 8.54]$.

Salary packages never lying in Negative Range so put '0'

Range will be _ [0, 8.5] with 90% Confidance.

Standard Normal Distribution ($\mu=0, \sigma=1$) is also called as z-distribution from t-score value we plot t-distribution. it is Similarly Normal Distribution.



i) z-distribution is taller than T-distribution, both the distributions lying on same mean ie ($\mu=0$)

ii) Both the distribution values are lying on 3 std away, in normal distribution it contains 99.7% of data.

iii) if we observe T-dist both the tail parts containing long tail as compare to normal distribution. by this there is chance for points to getting more far away from mean, area under the distribution-z is very less as compare to T-distribution. Cause of long tail

By the above mentioned reasons we getting error in std $\frac{\sigma}{\sqrt{4}}$ → it is called as Standard error.

By standard error we getting long range interval b/w two std values in above example we got (0, 8.5)

Hypothesis Testing :-

Opposite H_0 → Null Hypothesis ⇒ Ground truth ⇒ $=, \geq, \leq$

Opposite H_1 → Alternate Hypothesis ⇒ Bold claim ⇒ $\neq, >, <$

Ex : Osmania College announce average package of fresh graduate from 2020 - 2021 batch is atleast 10 LPA.

In this case their Statement Consider as H_1 it is a bold claim. So they said atleast 10 LPA So their may be chance of getting atmost Salary (ie above 10 LPA) or equal to 10 LPA. So we are formulate that as graterthan or equal ($\geq$) but in Alternate Hypothesis we dont have this Operator So we have to use ($>$) Symbol.

$$H_1 : \mu_{package} > 10 \text{ LPA}$$

So now we have to take Ground truth (ie H_0) value it is opposite to alternate hypo (H_1) ie —

$$H_0 : \mu_{package} \leq 10 \text{ LPA}$$

Step - 1 → formulate the $[H_0, H_1]$

Step - 2 → Collecting the samples (Ex - 3.5, 6, 2.5, 3.5, 3.7) and Calculating the mean of Sample. $\bar{x} = 3.84$.

Step - 3 → We have to appling Statistical operations. in that our Bold claim (H_1) is not Similar or promnising value in that Observation. So in final step we concluded it as

Rejected Hypothesis (ie H_1 is rejected), if H_1 rejected ie H_0 is accepted. in statistical terminology. Instead of 'accept' we say we fail to reject null Hypothesis (H_0)

Ex-2. if we take Orders from zomato and we say the packet of chicken biryani not contain 500gms in this Scenario our claim is to be (H_1) ie alternative Hypothesis H_1 : $\mu \neq 500$ ie. (H_0 - Ground truth is $\mu = 500$) then we Orders Some Samples of biryani from zomato and that Contain [505gms, 500gms, 490gm, 495gms, 496gms] if we calculate the mean of Sample ~~X~~ if we ~~take~~ Consider two Cases -

i) $\overline{X}_1 = 498$

ii) $\overline{X}_2 = 450$

Here - $H_1 \rightarrow \mu \neq 500$; $H_0 \rightarrow \mu = 500$

but if we Observe 2 cases ($\overline{X}_1, \overline{X}_2$), the $\overline{X}_1$ more near to 500 as - Compare $\overline{X}_2$, So Simply we can say $\overline{X}_1$ near to null Hypothesis

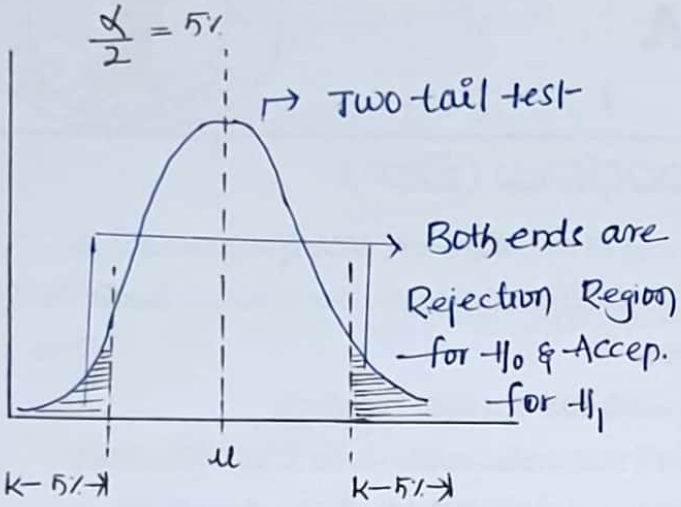
ie $\overline{X}_1 = 498 \Rightarrow$ fail to reject H_0 (or) Reject H_1

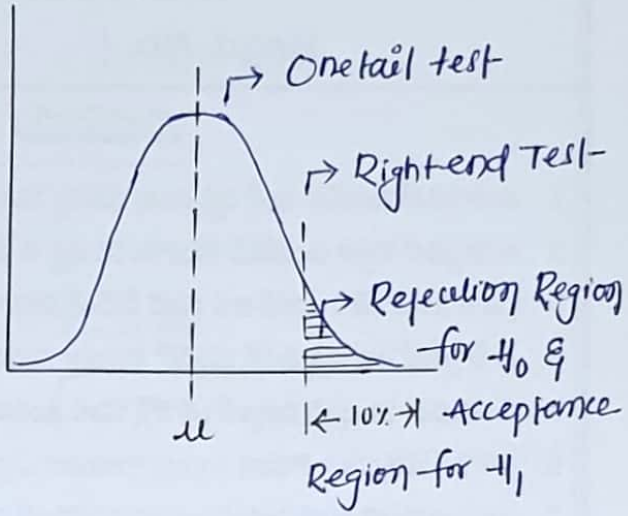
$\overline{X}_2 = 450 \Rightarrow$ Reject H_0 (or) Accept H_1

when ever $\neq$ Contain in H_1 after Calculating Significance Level this will going to two tail test if H_1 Contain $>$ then do right test if H_1 Contain $<$ then do left test

if Our Significance level $\alpha = 10\%$ and H_1 contain $\neq$

$\alpha = 10\%$ $H_1 \Rightarrow \neq$ $\alpha = 10\%$ $H_1 \Rightarrow >$





In

Significance level (α) :- it measures the Strength of evidance of a point that must be present in your Sample before rejeut Null Hypothesis (H_0) Simply Rejeution Region of a parameter.

Confidance Interval :- Estimates the Range of an Unknown parameter with Some Confidance level.

Confidance level :- it is the value, which Shows the probability of estimated location of parameter

Critical value

~~degree of freedom~~ :- Significance level/2 $\Rightarrow \frac{\alpha}{2}$

it is a value which represents lower bound, Upper bound ~~of~~ for Confidance Interval.

Degree of freedom :- $[t_{n-1}]$ it represent maximum no of logically Independent values.

Understanding Hypothesis Testing :-

1) 1) Alternate Hypothesis $H_1 \Rightarrow >, <, \neq$ (Bold claim)

2) Null Hypothesis $H_0 \Rightarrow \leq, \geq, =$ (Ground truth)

2) Collecting sample size n

Compute the mean from this sample $\bar{x}$

3) Compute the statistics

if population variance known $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

- z → Z-Score
- $\bar{x}$ → Mean of Sample
- μ → Actual mean (H_0)
- σ → std of population
- $\sqrt{n}$ → Sample size

if population variance unknown $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

- s → Std. of Sample
- t → T-score

4) Decide Significance level (α) means you need to stronger evidence to reject null-hypothesis, here Significance level tells us rejection region for Null-Hypothesis (H_0)

Hypothesis Testing is an act in statistics where test an assumption regarding a popular parameter and it is used to access the plausibility of a hypothesis by using sample data.

Difference b/w Parametric & Non-parametric test :-

Parametric tests are based on assumptions about the distribution of population from which the sample was taken.

Non-parametric test are not based on assumptions, that the data is collected from a sample that doesn't follow specific distribution.

Parametric test assumes a normal distribution of values or a bell-shape curve, where Non-parametric tests are used in cases where parametric tests are not appropriate.

Z-test, T-test, F-test and Anova test are parametric tests.

where chi-square is a non-parametric also called as distribution-free test In this the measurement of all the variables ie nominal or ordinal.

T-Test :- $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$

* T test are used when sample size is small ($n < 30$)
* and we have to caliculate degree of freedom ($n-1$)
* T-test used for test of significance of regression coefficent in the regression model.
* we use T-statistics when parameter of population are normal.
* population variable are unknown.
* Correlation coefficent in population is zero. then used t-test.

Note :- if we consider one continious variable

Then we use T-test.

—:

Z-Test :- $Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

* When population coefficient correlation is not zero then we use z test. In this sample size is large ($n > 30$) then concider.
* z-test used to determind whether 2 population mean are diffrent when population varience is known.
* z-test based on standard Normal distribution. and also called as large sample test.

F-Test :- (varience ratio - Test)

* F-Test is also called as ANNOVA Test (one-type of)
* F-Test is used to the two Independent estimation of population varience
* Two samples have same varience ie $[S_1^2 \& S_2^2]$
* F-Test is a small sample test.

$$F\text{-value} = \frac{\text{large sample varience}}{\text{small sample varience}} = \frac{S_1^2}{S_2^2}$$

* Degree of freedom for large population varience is V_1 and smaller is to V_2
* The Null Hypothesis of two population varience are equal to

$$H_0 = S_1^2 = S_2^2$$

degree of freedom for V_1 (larger) $= n-2$

degree of freedom for V_2 (smaller) $= n-1$

* F-Tests are used by comparing the ratio of two variances.
* The sample must be independent.
* F-Test Never be negative (-ve) because upper value is always grater than the lower value [S_1^2 (larger) / S_2^2 (smaller)]

Difference b/w T-Test, Z-Test, F-Test :-

	T-Test	Z-Test	F-Test
①	for Small Sample	large Sample	Small Sample
②	Population Coefficent Correlation is zero	Population Correlation Coefficent is not zero	Two Independent Estimation of Population.
③	Varience is Unknown	Varience known	Same varience
④	In multiple regression with '3' (Three) Individuals.		Testing for Overall Significance.

Non-parametric Test [chi-square] :-

* chi-Square denoted as χ^2, it is a Sampling analysis for testing Significance of population varience
* Due to Non-parametric test, it can be used for test of Goodness of fit ⇒ ie [dependent or independent], prefer for test of Independence
* this uses Simple random Sampling method.
* chi-Square test values lies in between 0 and 1 mostly it appears like lognormal distribution, mostly used for variables variables Independent test.

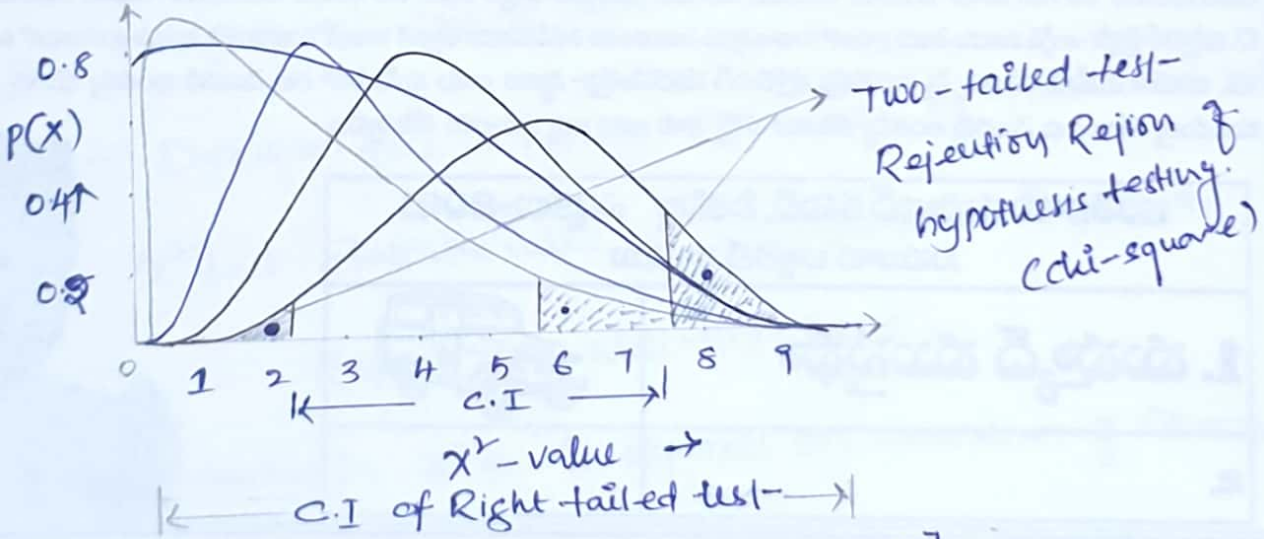
Chi - Square Test (χ^2) :-

It is a hypothesis testing method used to compare observed results with expected results, the purpose of this test is to determine if a difference between observed data and expected data is due to chance, or it is due to relationship between the variables which we are studying.

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i}$$

Where — χ^2 - chi square
O_i - observed value
E_i - Expected value

When we plot the diagram for check relationship between variables (ie whether dependent or Indipendent) mostly looks like lognormal distribution or pareto distribution mostly, it didn't contain negative values.



Different types of plots of [chi-square (χ^2)]

Conditions for Using chi-Square (χ^2) test :-

1) Total frequency (sample size) is large ($n > 50$)

2) Samples are independent (to check Indiperdency we use hypothesis testing where test condition is chi-square test (χ^2-test) ie [χ^2-rule] insteared of [z-scr or T-scr].

3) Cell freequency are linear.

Procedure for chi-square (χ^2) test of Independence by (Hypothesis)

Test Using :-

Step-1 :- Deside the Null Hypothesis (H_0) (Ground Truth) $= \geq \leq$

Alternate Hypothesis (H_1) (Bold claim) $> < \neq$

Step-2 :- Sampling { Collecting sample size (n)

Test Statistics value $= \sum \frac{(O_i - E_i)^2}{E_i}$

Step-3 :- Test statistic χ^2- score [Test of Independence]

Step-4 :- Deciding Significance level (α) by this we get or compute χ^2-Critical value.

Step-5 :- Decision Rule → plotting distribution by formulated value

if χ^2 test Statistic value > χ^2-critical ⇒ Reject H_0

∴ Reject our Null Hypothesis.

Degree of freedom in χ^2 ⇒ it depends on number of Columns

for Suppose dataset Contain '2' columns then - df = $(c_1 - 1) * (c_2 - 1)$

values present in — column-1, column-2

for Chi-Square (χ^2) we need Observed value and expected values for that calculated by -

① Observed value — For this we have to look into Observed frequencies in given sample (ie dataset)

② Computed Expected values by. (under Null Hypothesis (H_0) assumption)

$$\text{Expected value} = \text{rowtotal} * \frac{\text{Column total}}{\text{grand total}}$$

after getting this two values we have to do test of χ^2,

Step-4 :- In this decide α, $df = (\text{rows}-1)(\text{Col}-1)$

Step-5 :- χ^2 - test : if $\chi^2 > \chi^2_{df,\alpha} \Rightarrow$ Reject H_0 (Null Hyp)

p value test : $p\text{-value} = (1.0 - \text{cdf}(\text{Test statistic}))$

↓

[Cumulative distribution function]

cdf is another method to describe the description of random variables.

Now if $(p\text{-value} < \alpha) \Rightarrow$ Reject Null Hypothesis (H_0)

or

Accept Alternate Hypo (H_1)

p test is just extra one more test for evidance of Strong for reject Null Hypothesis (or) Accept (H_1) Hypothesis after χ^2 - test

Note :- If we have two categorical features, and trying to come to Conclusion Then we used chi-Square (χ^2) test-

CRISP-DM Framework