

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

Measure of Centre is Statistical measurement and it represents Summary of Data. i.e. 'Measure of Central Tendency' \rightarrow Distribution of Data ('column')

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

Mean :- Measure of $\frac{\text{avg. of all the values in Sample}}{\text{Sample}}$ is called mean

Represented by M $\Rightarrow M = \frac{\sum_{i=1}^n u_i}{n}$ \rightarrow NO of observations.

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

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

Ex: $[10, 20, 30, 40, 50] \Rightarrow$ '5' NO of Observations. $= \frac{5+1}{2} = \frac{6}{2} = 3^{\text{rd}}$

3^{rd} position value is - '30' \Rightarrow Median.

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

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

Dif blw mean & median :- * Both are measure the central Tendency

mean \rightarrow Gets Impacted with the presence of Outliers

median \rightarrow Won't Get Impacted with Outliers.

mean \rightarrow Time Complexity $O(n)$ less for mean.

median \rightarrow Time Complexity $O(n \log n)$ more for median.

	NO outliers	Outliers
Mean	—	\times
Median	—	\times

\times } Time Complexity

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

(ie) mostly occurring element in dataset

Ex : $[1, 2, 3, 4, 4, 3, 4, 2, 5, 7] \rightarrow$ So at Step (1)

$[1, 2, 2, 3, 3, \underline{4, 4, 4}, 5, 7] \rightarrow$ mode = 4 - Uni mode

$[1, \underline{2, 2, 2}, 3, 4, \underline{5, 5, 5}, 6, 6, 7] \rightarrow$ mode = (3, 5) - Bi mode

$[1, 2, 3, 4, \underline{5, 5}, \underline{6, 6}, 7, \underline{8, 8}] \rightarrow$ mode = (5, 6, 8) - Tri mode

Mean — When outliers present — ✓ [Affected]

Median — When " " — ✗ [Not Affected]

Mode — " outliers " — ✗ [Not Affected]

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

We have (4) measures — ① Range — $\text{Max} - \text{Min}$.

② Inter Quartile Range.

③ Variance

④ Standard Deviation.

Ex : $[10, 20, 30, 40, 50] = [50 - 10] \Rightarrow 40$. (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, mult also)

↳ 1) Measure of central tendency → 1) Mean (u)

2) median

3) mode

2) Measure of Spread.

→ 1) Range ② IQR ③ Variance ④ Std. (6)

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

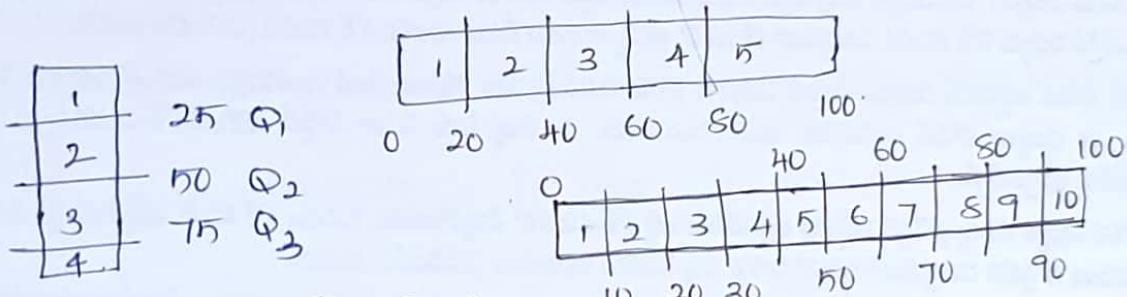
Ex: If we attend an exam of 100 marks and we got 65 marks in it $[65/100]$ and ranked as 1st then - 100% percentile. It tells us no one or any value above the mark that i got. So here my score is 100th percentile.

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 divided in to 4 equal no of parts.

Quintile mean " " 5 equal no of parts

Decile mean " " 10 equal no of parts



$$\text{IQR} = \text{Median} = Q_3 - Q_1 \Rightarrow Q_2$$

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

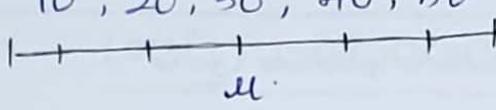
$$\text{if we want } \text{IQR} = Q_3 - Q_1 = 75 - 25 = 50 \Rightarrow Q_2$$

Simply we can say IQR is = 50th percentile.

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

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

Ex: $10, 20, 30, 40, 50 \Rightarrow \text{mean} = 30 = \bar{m}$.



Variance represents in (σ^2) .

$$\sigma^2 = \frac{\sum_{i=1}^n |\bar{m} - \text{Obj}_i|}{n}$$

$$\begin{aligned} 30-10 &= 20 \\ 30-20 &= 10 \\ 30-30 &= 0 \\ 30-40 &= -10 \\ 30-50 &= -20 \end{aligned} \quad \left. \begin{array}{l} \\ \\ \\ \\ \end{array} \right\} \text{points.}$$

Standard Deviation : is a square root of ~~standard deviation~~ Variance

Denoted as $(\sigma) = \sqrt{\text{Variance}}$.

M. of Spread:

- Range - when outliers imputed - ✓ (Affected)
- IQR - when outlier " - ✗ (Not Affected)
- Variance - " outliers " - ✓ (Affected)
- Std. - " outliers - ✓ (Affected)
- MAD - mean absolute Deviation - " - ✗ (Not E)

Mean 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.

$$\text{M.A.D} = \text{median} |\bar{m} - \text{Obj}_i|$$

In this it rejects all outliers

* Standard Deviation gives Distribution of variables.

By above all topics we covered on ~~on~~ Uni variables which perform

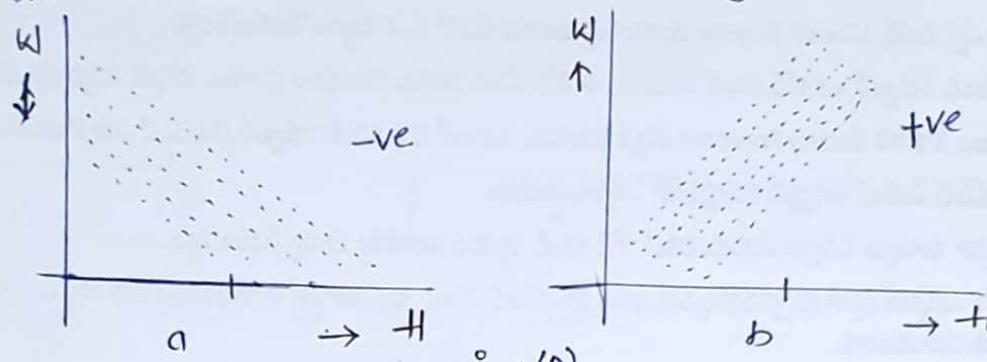
Uni variable $\begin{cases} \text{Measure of Spread} \\ \text{Measure of Central Tendency} \end{cases}$

M. Relation :-

On Bi variables we find $\begin{cases} \text{Co-variance} \\ \text{Co-Relation} \end{cases}$ } measure of relation.

* Univariable ie deals only one column, or one variable where Bi variable means deal on two variables or two columns.

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



In fig a when height ~~increase decrease~~ weight ~~also decrease~~ increase this tells the Correlation b/w two variables ie -ve Correlation ie negative cov

if fig b observe height when increase weight also increase this tells

relation b/w two variables is positive ie Covariance is positive.

If we observe Variance of Univariable $\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$

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

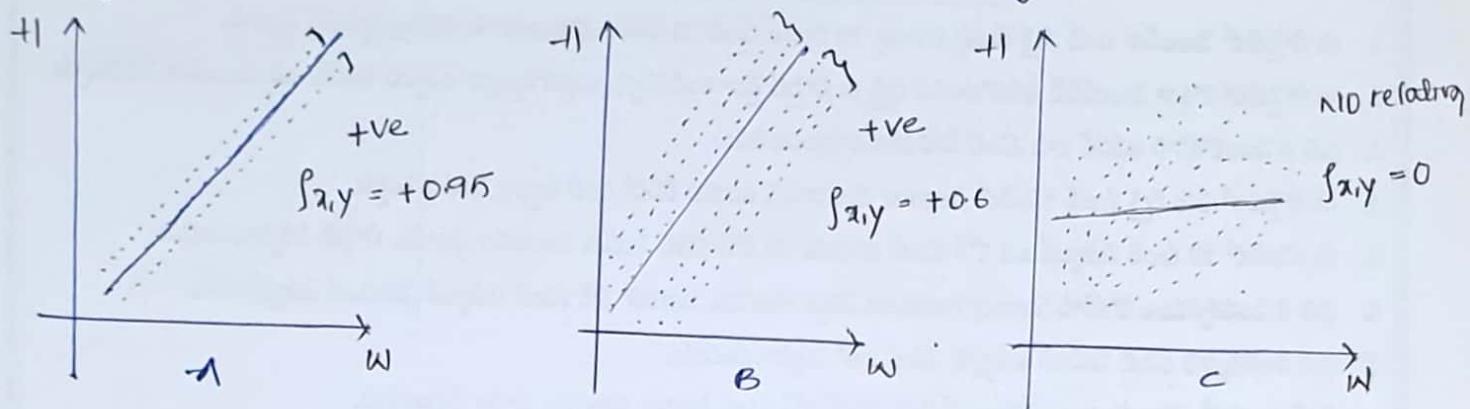
which tell the $\frac{\text{sum of}}{\text{diff b/w mean - to each data point in x axis and mean on 'y' variable - to each point}}$

Correlation :- it measures the degree to which two variables move in relation to each other in other it call as Pearson Correlation Coefficient and denoted as $r_{x,y}$ \therefore

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

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

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



Above figures tell us there is positive correlation 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 that -ve correlation.

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

Diff b/w covariance and correlation :-

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

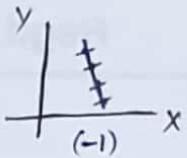
Correlation on the other hand measures both the strength and direction of the linear relationship b/w two variables

Correlation coefficient Support linear relationship.

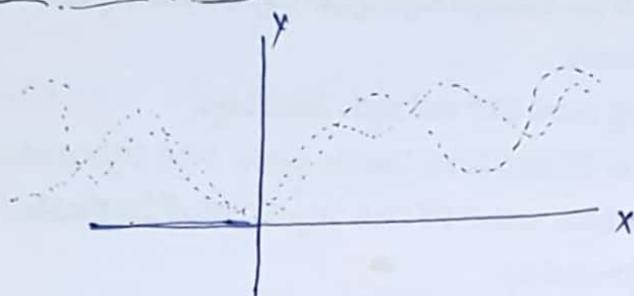
if we observe figures A, B. both tell (+ve) correlation but strength of the relation is changed, cov doesn't tell strength

Both tells us direction, linear relationship of two variables.

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



Problem with Pearson Correlation Coefficient :-



→ Bimodal data. Varies like this

P. Correlation Coefficient only tell Linear Relationship. won't tell Non-linear R. In. L.R only, $\sum_{i=1}^n x_i y_i = 0$ values in otherwise it fails ie $\sum x_i y_i \neq 0$

Probability :- it is Study of Uncertainty.

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

Ex: Tossing a coin, Rolling a die, Picking an object

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

Ex: Tossing a coin — $\{H, T\} \Rightarrow$ S.S.

Rolling Die — $\{1, 2, 3, 4, 5, 6\}$

Event :- Any Subset of Sample Space

Ex: Tossing TWO COINS. \Rightarrow Random Experiment

SS $\Rightarrow \{HH, HT, TH, TT\}$

Event $\Rightarrow E_1 = \{HT, TH\}$ — Getting Exactly 1 Head $\Rightarrow \frac{1}{2}$

$E_2 = \{HH, TT\}$ — Getting something on Both Toss $\Rightarrow \frac{1}{2}$

$E_3 = \{HH\} \rightarrow$ Getting Two heads $\Rightarrow \frac{1}{4}$

$E_4 = \{TT\} \rightarrow$ " Two Tails $\Rightarrow \frac{1}{4}$

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

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

Rules / Axioms of Probability :-

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

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

③ Probability of Events that are mutually Exclusive

for any sequence of 'Events' that are mutually Exclusive

$$\text{Ex: } S_1 = \{1, 2, 3\}$$

$$S_1, S_2 = \left\{ \begin{array}{c} 1 \\ 2 \end{array} \right\} \cap \left\{ \begin{array}{c} 3 \\ 4 \\ 5 \end{array} \right\}$$

common value = 3

$$S_2 = \{3, 4, 5\}$$

$$S_1, S_3 = \left\{ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \right\} \cap \left\{ \begin{array}{c} 4 \\ 5 \\ 6 \end{array} \right\}$$

NOT mutually Exclusive

$$S_3 = \{4, 5, 6\}$$

$$S_2, S_3 = \left\{ \begin{array}{c} 4 \\ 5 \\ 6 \end{array} \right\}$$

M. Exclusive

By the above we can say —

$$E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n = \emptyset \rightarrow \text{M. Exclusive}$$

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

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

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

Events like — $E_2 = \{TT\} = \frac{1}{4}$

$E_3 = \{HH, TT\} = P(E_3) = \frac{1}{2},$

if we observe - Event E_3 $P(E_3)$ that belongs Union $E_1 \cup E_2$

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

$$P(E_3) = \boxed{P(E_1 \cup E_2) = P(E_1) + P(E_2)} \quad \text{if } \downarrow P(E_1 \cap E_2) = \emptyset$$

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

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

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

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)

$$\text{S.S} = \{1, 2, 3, 4, 5, 6\}$$

Event — Getting prime no $\{2, 3, 5\}$

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

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.

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

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

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

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

All possible outcomes = 1 Outing

$$\text{Sample Space} = 6 \Rightarrow P(A \cap B) = \frac{1}{6}$$

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

(10) (Q1) - $P(A|B)$ - Conditional prob - $P(\text{Getting prime} | \text{Even no})$

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

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

CRISP-DM FRAMEWORK :-

Cross Industry Standard process for data mining.

① Business Understanding \rightarrow 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, Visualization.

③ Data Preparation \rightarrow Treat outliers, missing values, encoding data, convert data.

④ Machine Learning modeling.

⑤ Evaluation of M.L. Model

⑥ Deployment

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 - $\{H, T\}$

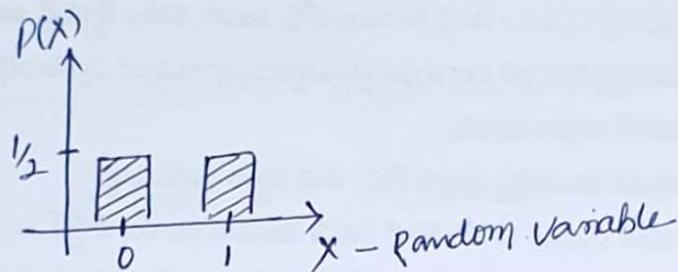
Random variable - 'X' - Counting the no of heads - $\{H, T\} - \{1, 0\}$

Computation - Compute the probability of 'X' = P.d.f

$$P(X=0) = P(\{T\}) = \frac{1}{2} \quad \begin{bmatrix} H \\ T \end{bmatrix} \rightarrow 1$$

$$P(X=1) = P(\{H\}) = \frac{1}{2} \quad \begin{bmatrix} H \\ T \end{bmatrix} \rightarrow 0$$

Plotting - Plot the probability distribution function (P.d.f)



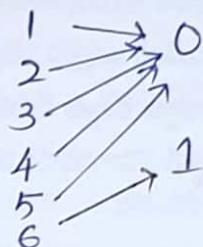
Ex-2 R.E - Rolling a dice

$$S.S = \{1, 2, 3, 4, 5, 6\}$$

R.V - Getting a '6' : X

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

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



Plotting -



* There is more chance of getting failure than success

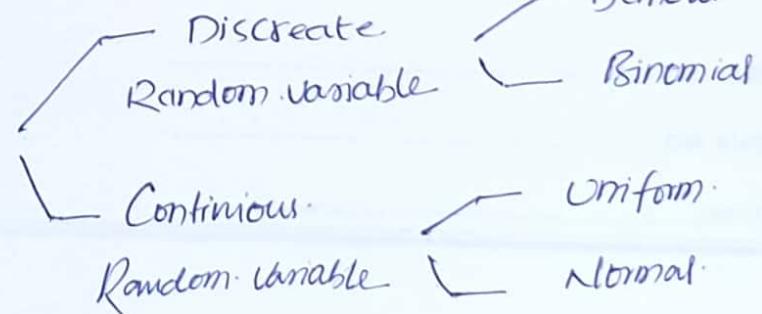
Simply we say Random variable is a function which will take Sample Space as input & give some O/p in a form of a number

$$X : \{H, T\} \rightarrow \{1, 0\} \Rightarrow O/p$$

↳ S.S.

Classification :-

Random variables



distplot \rightarrow distribution plot it consist of three things.

- ① K.de \rightarrow pdf
 - ② Histogram
 - ③ cdf
- This function distplot is discontinued in the future we have to use a new function called as "displot"

Discrete Random Variable :- It has a countable no of possible values. The Probability of each value of a discrete random variable is between 0 and 1 and Sum of the probabilities equal to 1.

Bernoulli Random variable :- if the outcome taken Countable values (finite) Set as well Random variables Containing two values only (ie Binary set of values) is called Bernoulli Random variable.

Ex: $\{0,1\} \Rightarrow$ Here '0' represents failure and '1' is Success. This variable contains only two values that are finale number and having only two values 1) failure, 2) Success. Such variable is called as "B.R.V" P.M.F \rightarrow Probability Mass function is used when we have discrete Random variables it tells us variable is Bernoulli or not

$$\begin{aligned} \text{P.M.F} \Rightarrow P(X=1) &= P & - & \text{Success} \\ P(X=0) &= 1-P & - & \text{Failure} \end{aligned}$$

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

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

$\begin{array}{c} / \backslash \\ S/F \quad S/F \end{array}$

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.

→ Shows Density of a variables.

$HH = 2 \quad \{HH, HT, TH, TT\} = \{2, 1, 0\}$

$\begin{array}{l} HT = 1 \\ TH = 1 \\ TT = 0 \end{array}$

$$P(X=0) = P(\{TT\}) = \frac{1}{4}$$

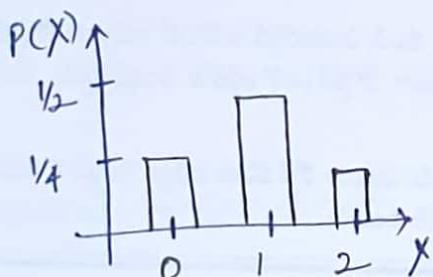
$$P(X=1) = P(\{HT, TH\}) = \frac{2}{4}$$

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

4. Compute the probability of R.V. $\Rightarrow P(X=0, 1, 2)$

5. Plot PDF

$\Rightarrow P(\{0, 1, 2\}) \rightarrow$ is not a Bernoulli



but it is a collection of Bernoulli.

\Rightarrow Simply called as Binomial.

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

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

i = no of Success

P = probability of Success

$1-P$ = probability of failure

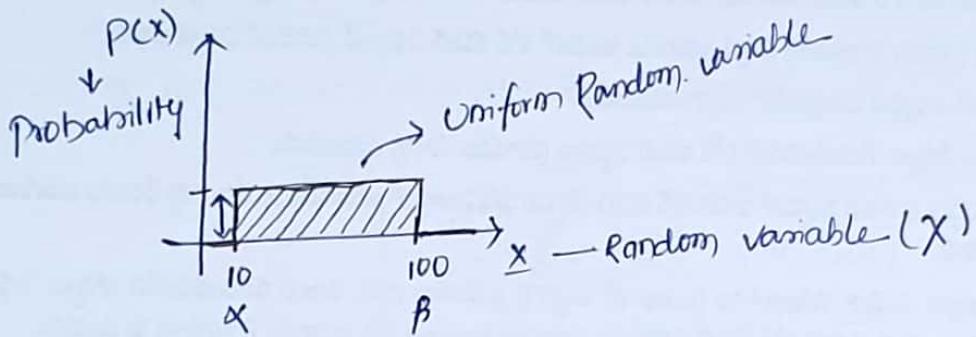
n = no of trials.

Continuous Random Variable :- Uniform, Normal

'X' is a C.R.V. if the set of possible values are infinite or Uncountable is called as Continuous Random Variable

Uniform Random Variables :-

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



$$P(X=x) = \begin{cases} \frac{1}{\beta-\alpha}, & \text{if } \alpha \leq x \leq \beta \\ 0, & \text{otherwise} \end{cases}$$

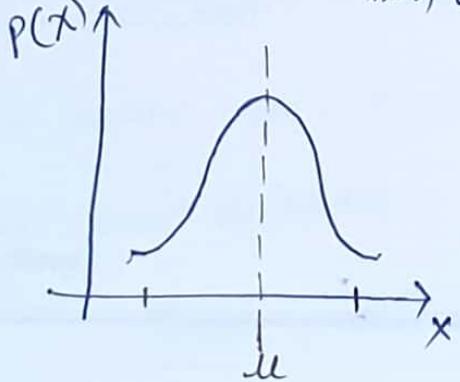
① R.V. x lies in between (α, β) where α is min value, β is max

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

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

(15)

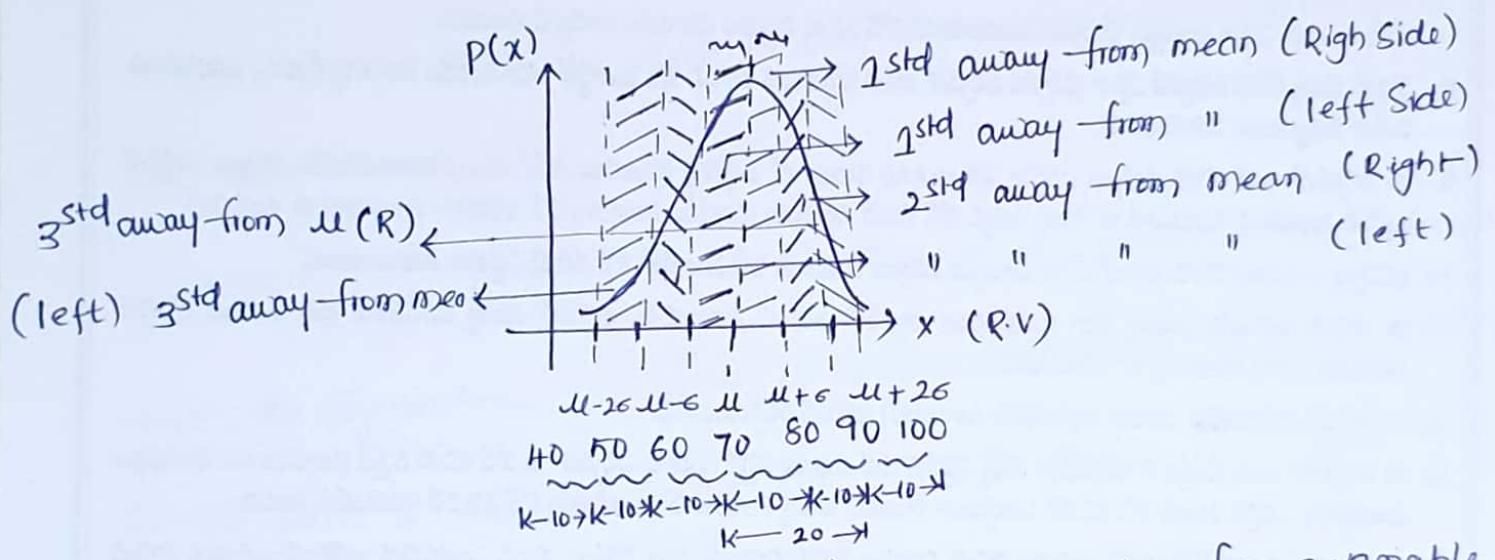
Normal Distribution :- This distribution also called as "Gaussian D."

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 rule $\rightarrow 68 - 95 - 99.7\%$

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

$$X \sim (\bar{X}, 100) \Rightarrow \bar{X} = 70, \sigma^2 = 100, \sigma = 10$$



Simply we can say that 1^{std} lies 68% of data from a variable

ie [60, 80] lies 68% of data.

2^{std} lies 95% of data from variable ie [50, 90] $\sim 95\%$

3^{std} lies 99.7% of data from variable ie [40, 100] $\sim 99.7\%$

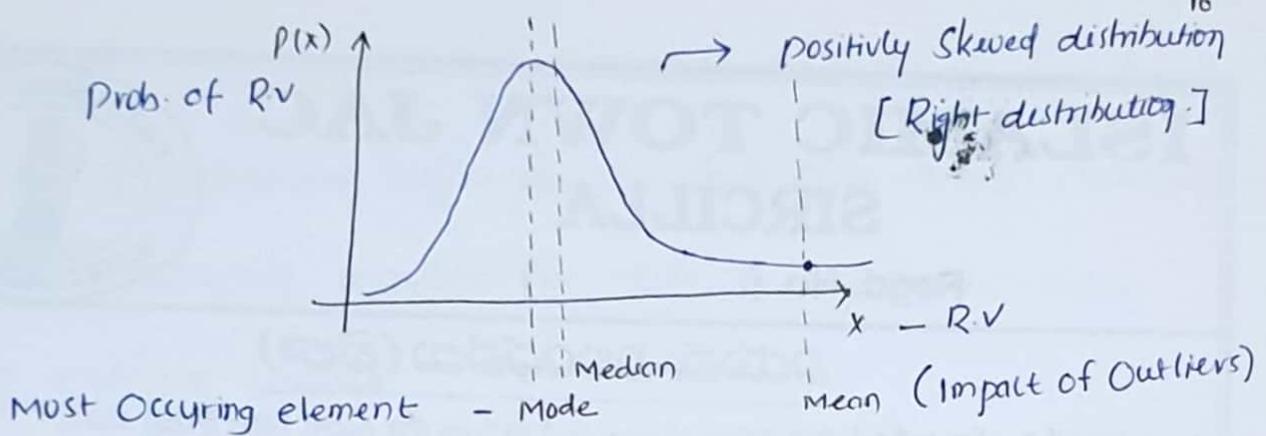
Q: What happens to mean median mode, when you have positively or negatively skewed distribution?

①

$$\text{Mode} \leq \text{median} < \text{mean}$$

\bar{X} 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 some 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 less than -0.5 or more than $+0.5$ it is not symmetric.

If skewness value lies in bw -1 to $+1$ than it is called as heavily negatively 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 look like peaked, if the value is negative its look like flat.

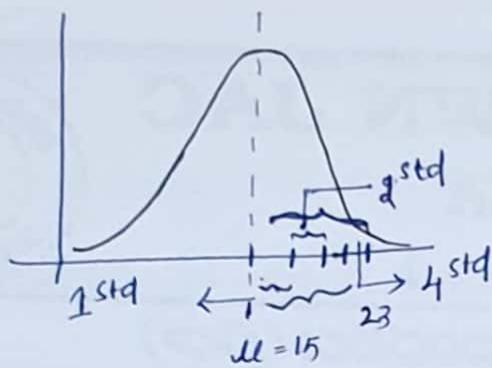
if it is
Standard Normal Distribution :- A Random variable \uparrow follow normal distribution with the mean of zero and standard deviation is one, then this type of variable is called as standard normal variable, and the distribution known as standard Normal distribution.

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

Random Variable \hookrightarrow Normal Distribution

Z-score :- it is a measure of how many std away ~~is~~ 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



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

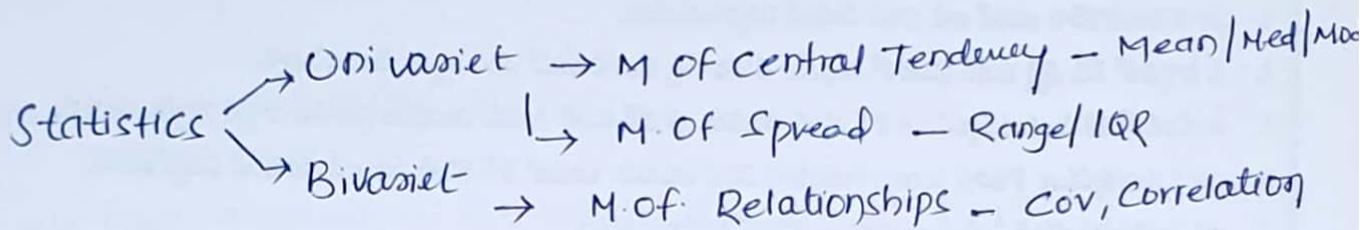
$$\sigma = 2$$

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

X_i - point in variable

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

Distributions & Transformations :-



Probability → Basics - R.E, S.S, Event

↳ Axioms of probability.

↳ Conditional probability

R.E.

S.S

R.V

↳ Random variable

Prob. distribution

↳ P.d - Discrete → PMF

Plotting. P.d.

↳ P.d - continuous → PDF

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)

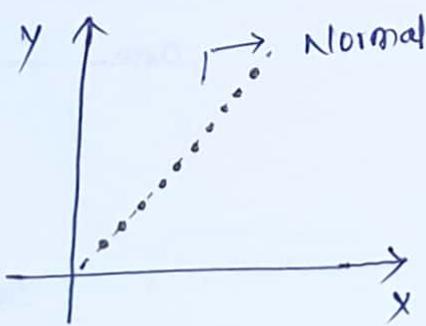


fig (1)

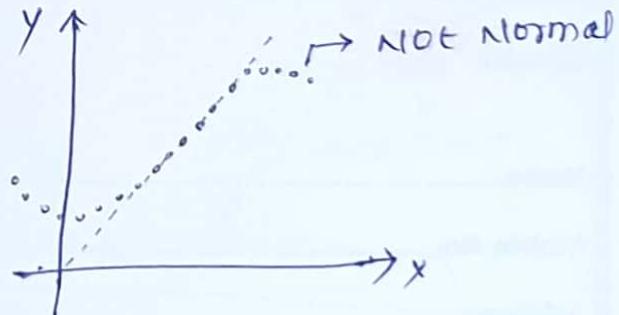


fig (2)

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

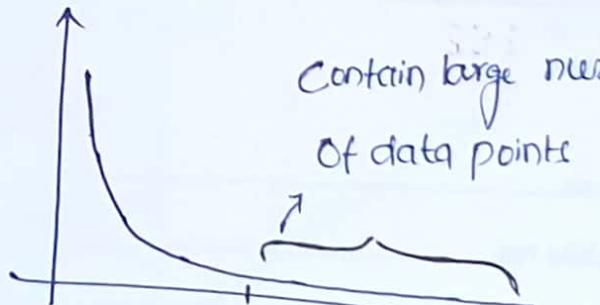
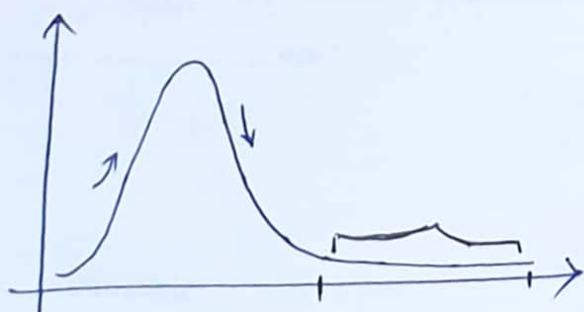
How to get a QQ plot :-

$$\mu=0, \sigma=1$$

- ① we have to check ① Theoretical Quantity - Normal distribution.
 - ② Observed Quantity - Given distribution / column.
 - ③ Sort the values of Th.Q. and Obs.Q.
 - ④ 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 :-

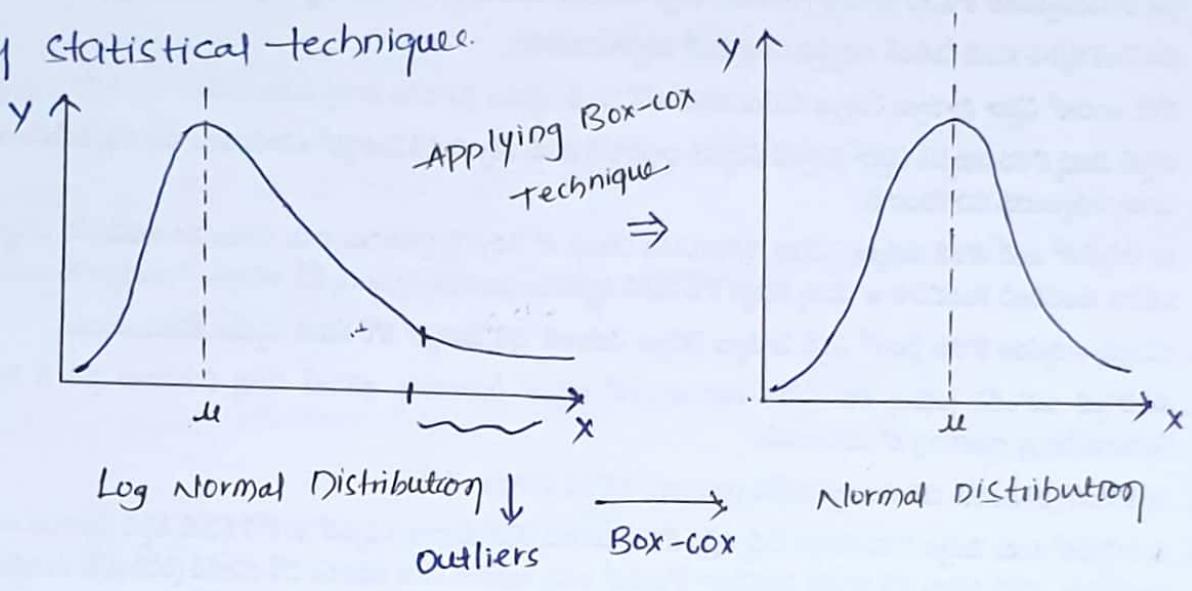
Pareto Distribution :-



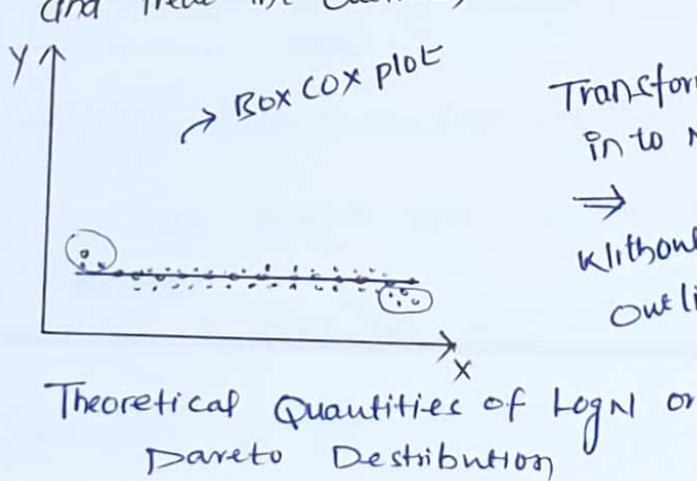
if we draw graph on Salary, houseprice, carprice etc columns we get this type of distribution.

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

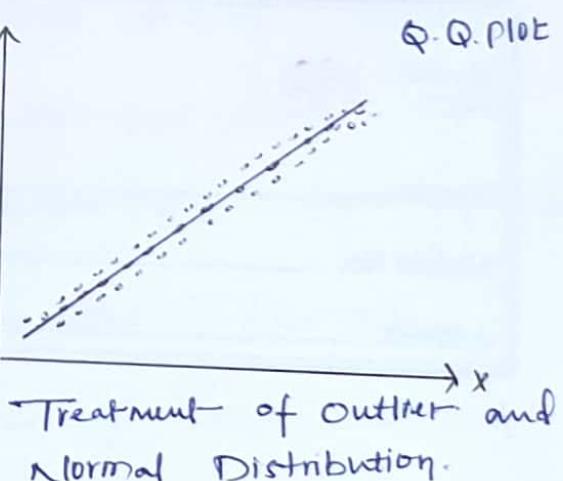
Box-Cox Transformation :- It is a transformation technique of a non-normal dependent variable like (Log Normal D, Pareto D) in to normal shape, Normality is an important assumption for many statistical techniques.



- Some more Distributions — Log Normal, Pareto
- Q.Q plot → Using for test of Normality.
- Box Cox → Transformation (Convert log-N, Pareto → Normal Distribution and treat the Outliers)



Transformed into Normal
→ Without remove
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 determine the Inferential Statistics.

Point Estimation :- if we want to find the mean of (u) population,

firstly we draw a Sample and then find Sample mean (\bar{x}). and ^{etc}

estimate the population mean ($u \approx \bar{x}$) 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 'u' and sample mean \bar{x}

$$u = \frac{\sum_{i=1}^N \text{Obs}_i}{N} ; \text{ Population Variance } \sigma^2 = \frac{\sum_{i=1}^N (\text{Obs}_i - u)^2}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^n \text{Obs}_i}{n} ; \text{ Sample Variance } s^2 = \frac{\sum_{i=1}^n (\text{Obs}_i - \bar{x})^2}{n-1}$$

Note: (n-1) Coarse of bias in U.R.S. (\therefore Remove Impact of bias)

We have diff types of Sampling Techniques Such as Convenient, Convenience

Sampling, Uniform random Sampling, in this Convenient and Convenience Can't give you perfect value of mean (ie \bar{x} Sample mean $\bar{x}_s = \text{Population } u_{pop}$)

In U.R.S is gives you some what approx mean value. It have also.

Some bias - for that we put (-1) in denominator of Sample variance

22

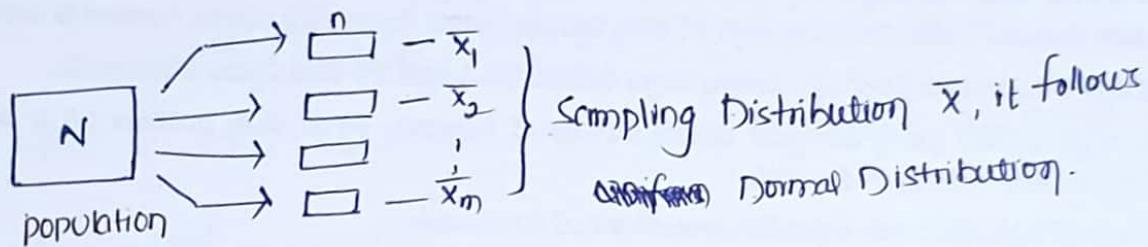
Central Limit Theorem :- If we have large population and we divided into many samples then the mean of all the samples from the population will be almost equal to mean of the entire population ie each of the sample normally distributed

$$\mu_{\bar{x}} \simeq \mu_{\text{pop.}} \rightarrow \text{Mean of population.}$$

↳ 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



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

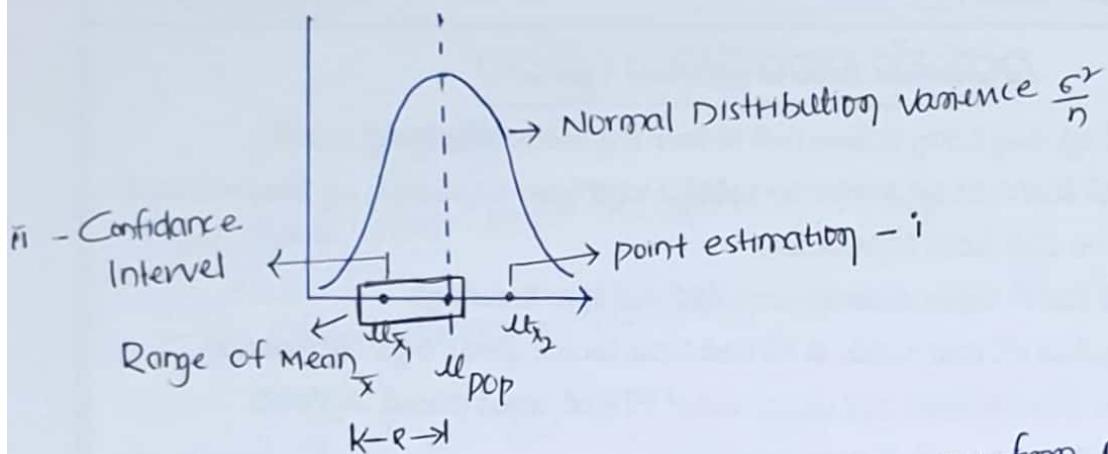
2) Sample D. Variance = $\frac{\sigma^2}{n}$ { n - size of each sample. σ^2 - pop variance. }

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

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

After Sampling distribution every sample mean [$\mu_{\bar{x}} = \mu_{\text{pop.}}$] mean of population So if we consider one sample from Sample distribution - the Sample can be biased or convenient/voluntary Sample and point estimates that's why we have to go Confidence Interval.

- 1) $\bar{u} \approx \bar{x} \rightarrow$ point Estimate, Sample can contain bias.
- 2) C interval $\rightarrow \bar{u} \approx [\bar{x}, \pm \text{Something}]$ with y% Confidence.



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

So that represent in Z^* -score value then -

std of pop is given

$$\bar{u} \approx [\bar{x}, \pm Z^* \frac{\sigma}{\sqrt{n}}] \text{ with y% of Confidence.}$$

if we find Z -score value then we easily represent how much % Confidence

if $Z^* \rightarrow 1$ std away from mean it represent - 68% of Confidence

Ex: Consider one college, in that Students got job 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 $\bar{u}_{\bar{x}} = 3.84$.

In this case we don't know the salary package of every student, who got job. So whatever salaries we are taken are Convenience Sample.

So $\bar{u}_{\bar{x}} = 3.84$ is point estimate, Biased. So we need to go Confidence

interval in this $[\bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}}]$ with y% Confidence, we don't know

the value of " σ_{pop} ", because \bar{u}_{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

$$S^2 = \frac{(3.84 - 3.5)^2 + (3.84 - 6)^2 + \dots + (3.84 - 37)^2}{5-1} \quad n=5$$

$$S^2 = 1.1586$$

if we have don't know the value of σ_{pop} instead of it we can use if we use it. z^* value also changed.

$S \rightarrow$ 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.

$$= \left[\bar{x} \pm t_{n-1, \alpha/2} \frac{S}{\sqrt{n}} \right] \quad \begin{matrix} \text{Sample Std.} \\ \downarrow \text{S.mean} \quad \downarrow \text{sample size} \end{matrix}$$

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

So here 90% can be represented as 0.9

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

Here 10% \rightarrow Significance Level (α)

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

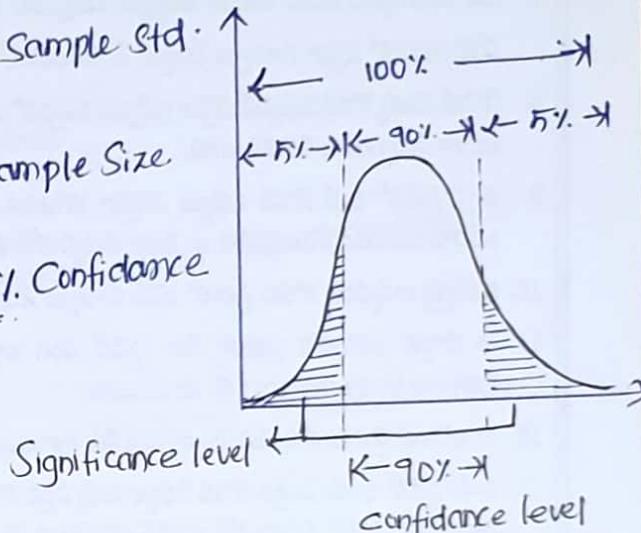
\downarrow Degrees of freedom \therefore Critical V = $\frac{\text{Significance}}{2}$

$$\text{DOF} = \text{Sample size} - 1. \quad ; \quad \text{C.V} = \frac{10\%}{2} = 5\% \approx 0.05$$

so $-t_{4, 0.05} \rightarrow$ we have to find the value from T-score

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

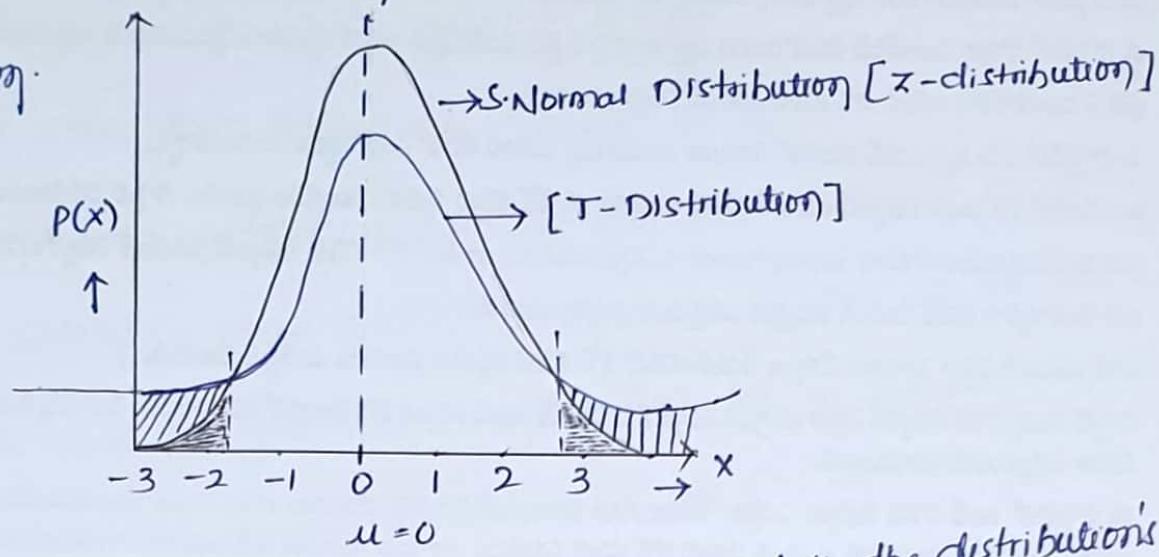
put all values - $\left[3.84 \pm 2.132 \times \frac{1.1586}{\sqrt{5}} \right]$ with 90% confidence
 $[3.84 \pm 4.7] \Rightarrow [-0.86, 8.54]$.



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

Range will be - $[0, 8.5]$ with 90% Confidence.

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 reason we getting error in std $\frac{\sigma}{\sqrt{n}} \rightarrow$ 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 $\leftarrow H_0 \rightarrow$ Null-Hypothesis \Rightarrow Ground truth $\Rightarrow =, \geq, \leq$
 $\rightarrow H_1 \rightarrow$ Alternate-Hypothesis \Rightarrow Bold claim $\Rightarrow \neq, >, <$

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 almost salary (ie above 10 LPA) or equal to 10 LPA. So we are formulate that as greater than or equal (\geq) but in alternate-hypothesis we dont have this operator so we have to use ($>$) symbol.

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

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

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

Step - 1 \rightarrow formulate the $[H_0, H_1]$

Step - 2 \rightarrow 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 \rightarrow we have to apply in Statistical operations. in that our Bold claim (H_1) is not similar or promising 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 500gme in this scenario our claim is to be (H_1) ie alternative hypothesis H_1 : $\mu \neq 500$ ie. (H_0 - Ground truth ie $\mu = 500$) then we Order Some Samples of biryani from zomato and that contain [505gms, 500gms, 490gm, 495gms, 496gms] if we calculate the mean of sample ~~will~~ if we ~~will~~ consider two Case -

$$i) \bar{x}_1 = 498$$

$$ii) \bar{x}_2 = 450$$

here - $H_1 \rightarrow \mu \neq 500$; $H_0 \rightarrow \mu = 500$.

but if we observe 2 cases (\bar{x}_1, \bar{x}_2) , the \bar{x}_1 more near to 500 as - compare \bar{x}_2 , so simply we can say \bar{x}_1 near to null hypothesis

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

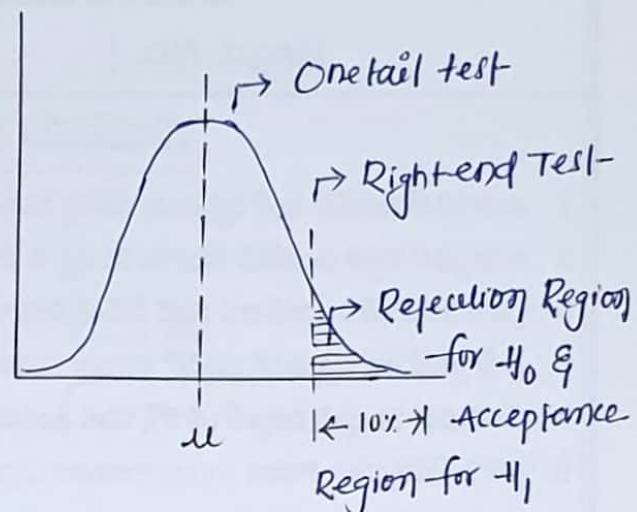
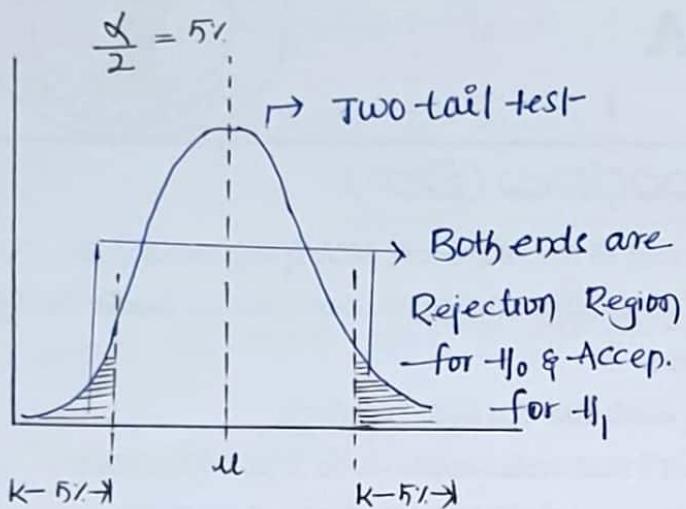
$\bar{x}_2 = 450 \Rightarrow$ Reject H_0 (or) Accept H_1

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 $H_1 \neq$

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

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



In Significance level (α) :- it measures the strength of evidence of a point that must be present in your sample before reject Null Hypothesis (H_0) Simply Rejection Region of a parameter.

Confidence Interval :- Estimates the Range of Unknown parameters with Some Confidence level.

Confidence level :- It is the value, which Shows the probability of estimated location of parameter

Critical value
Degree of freedom :- $\text{Significance level } \frac{\alpha}{2} \Rightarrow \frac{\alpha}{2}$

It is a value which represents lower bound, upper bound of for Confidence 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)
- 3) Collecting sample size n

Compute the mean from this Sample \bar{X}

- 3) Compute the Statistics

if population variance known

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \begin{array}{l} \text{Z-Score} \\ \text{Mean of Sample} \\ \bar{X} - \mu \rightarrow \text{Actual mean } (H_0) \\ \frac{\sigma}{\sqrt{n}} \rightarrow \text{std of population} \\ \sqrt{n} \rightarrow \text{Sample size} \end{array}$$

if population variance unknown $t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \rightarrow \begin{array}{l} \text{std. of Sample} \\ \frac{s}{\sqrt{n}} \rightarrow \text{T-Score} \end{array}$

- 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 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 test's.

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

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

* T-test are used when Sample size is small ($n < 30$)

* and we have to calculate degree of freedom ($n-1$)

* T-test used for test of significance of regression coefficient in the regression model.

* we use T-statistics when parameter of population are normal.

* population variable are unknown.

* Correlation coefficient in population is zero. then used t-test.

Note :- if we consider one continuous variable (one) then we can use t-test.

Then we use T-test.

:- nonparametric

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

31

- * When population coefficient correlation is not zero then we use \bar{Z} test. In this sample size is large ($n > 30$) then consider.
- * Z -Test used to determine whether 2 population mean are different when population variance is known.
- * Z -Test based on standard normal distribution. and also called as large sample test.

F-Test :- (variance ratio - Test)

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

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

- * Degree of freedom for large population variance is v_1 and smaller

is to v_2

- * The null hypothesis of two population variance 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 greater 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 Coefficient Correlation is zero	Population Correlation Coefficient is not zero	Two Independent Estimation of Population.
③ Variance is unknown	Variance known	Same Variance
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 variance.
- * Due to non-parametric test, it can be used for test of Goodness of fit \rightarrow i.e. [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.

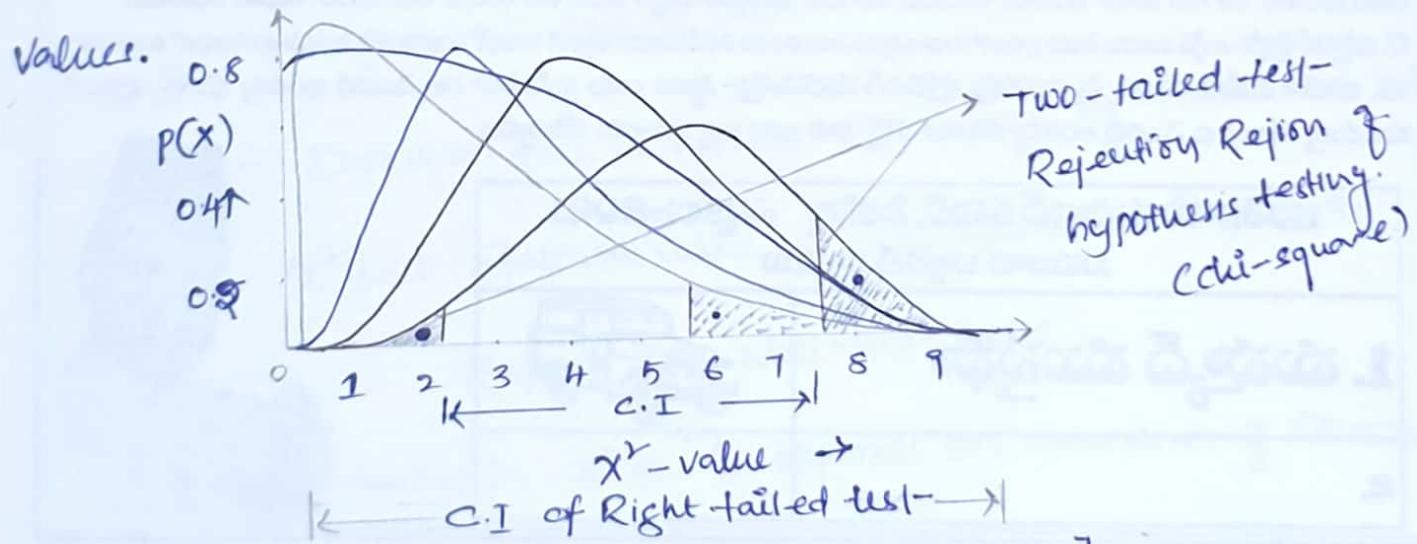
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where - χ^2 - Chi-square

O_i - Observed value

E_i - Expected value

When we plot the diagram for check relationship between variables (i.e. whether dependent or independent) mostly looks like lognormal distribution or parito 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 :-

34

- 1) Total frequency (sample size) is large ($n > 50$)
- 2) Samples are independent (to check Independence we use hypothesis testing where test condition is Chi-Square test (χ^2 -test) ie [χ^2 -rule] instead of [Z -scr or T -scr].
- 3) cell frequency are linear.

Procedure for Chi-Square (χ^2) test of Independence by (Hypothesis)

Test Using :-

Step-1 :- Decide the Null Hypothesis (H_0) (Ground Truth) $\Rightarrow \leq$
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 \rightarrow plotting distribution. by formulated value

if χ^2 test Statistic value $>$ χ^2 -critical \Rightarrow Reject H_0

\therefore Reject our Null Hypothesis.

Degree of freedom in $\chi^2 \Rightarrow$ 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 35 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)

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

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

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

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

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

or

Accept Alternative Hypothesis (H_1)

P test is just extra one more test for evidence of

Strong for reject Null Hypothesis (H_0) 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

