We can not Belive ... 0.05 Table

3

MT-3

Central limit Theorem for (Population) praportion:

Sampling Distribution of proportion's $\bar{P}$ For a Large Sample follows approximately normal distribution with P (The population proportion). And standard deviation $\sqrt{\frac{P(1-P)}{n}}$

Hypothesis Testing For Population Praportion:

→ Z-Test For Praportion:

According to Central limit Theorem of praportions the Sampling distribution of praportion $\bar{P}$ For Large Sample follows an approximate normal distribution then Z-statistic is

$$Z = \frac{\bar{P} - P}{\sqrt{\frac{P(1-P)}{n}}}$$

To calculate standard deviation $\sqrt{\frac{P(1-P)}{n}}$ we need knowledge of 'p' however we can use the value of $\bar{p}$ estimated from large samples.

One of the rule used is.

→ That the value of the $n \times \bar{p} \times (1-\bar{p}) \geq 10$. to use

$$= \frac{(\bar{p}-p)}{\sqrt{\frac{P(1-P)}{n}}}$$

① According to a study Exactly 12% Gift card's purchased from e-commerce how are never used. the Manager an e-commerce company wanted to Test wheather this claim is true. collected data 250 Gift card purchases and found the 22 Gift card's for hot used till it's expiry date.

$H_0: P=12$

$H_1: P \neq 12$

$n = 250$

$\bar{p} = 22/250 = 0.088$

$P = \frac{12}{100}$

$P = 0.12$

$n \times \bar{p} \times (1-\bar{p}) \geq 10$

$250 \times 0.088 (1-0.088) \geq 10$

$= 22 \times 0.912 = 20.064.$

$$z = \frac{\bar{p}-p}{\sqrt{\frac{P(1-P)}{n}}} = \frac{0.088-0.12}{\sqrt{\frac{0.12(1-0.12)}{250}}}$$

$$= \frac{0.88+0.12}{} - \frac{0.88-0.12}{}$$

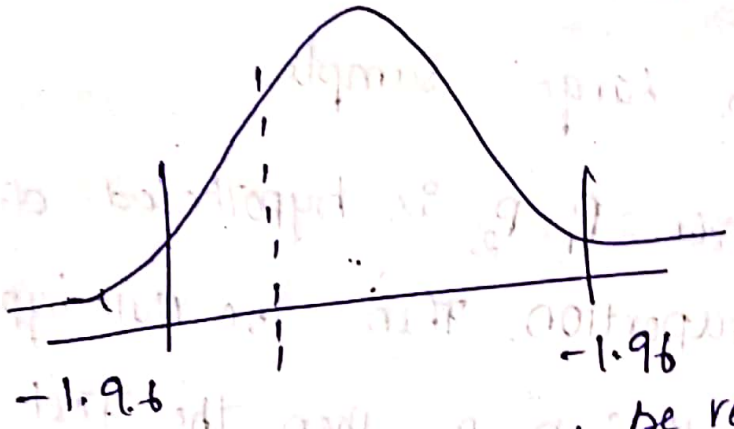
$$= \frac{0.088 - 0.12}{0.02055} = \frac{-0.032}{0.02055}$$

$$= -1.557.$$

$\alpha = 0.05$

$z = 1.96.$

It is two tailed Test



H_0 ~~is~~ accepted. can not be rejected.

confidence Intreval $\bar{x} \pm z \sigma/\sqrt{n}$

$$\bar{p} \pm 1.96 \sqrt{\frac{p(1-p)}{n}}$$

$$0.088 \pm 1.96 \left(\sqrt{\frac{0.12(1-0.12)}{250}} \right)$$

$$[0.0528, \quad 0.1231]$$

18/09/2018

Hypothesis Test for difference Population praportion, TWO Sample z-Test of Praportions

→ When the praportions estimated from Large sample then sampling distribution of praportions follow a normal distribution According to Central limit Theorem.

Let $\bar{P}_1$ & $\bar{P}_2$ be estimated values of praportion from Large sample.

The Difference $P_1 - P_2$ is hypothesed difference in population praportion. Then the Null hypothesis is $P_1 = P_2$ (i.e $H_0: P_1 = P_2$ then the Test-statistic is

$$z = \frac{(\bar{P}_1 - \bar{P}_2) - 0}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where $\bar{P} = \frac{n_1\bar{P}_1 + n_2\bar{P}_2}{n_1 + n_2}$ is the pooled estimate for praportion.

(1) The Marketing Manager for a company belives that non-affluent Customers are Sensitive to discount Compare to affluent Customers to valideti this hypothesis discount coupons are send to non-affluent and affluent Customers and the data is provided below.

Group	Sample Size	no of customer using coupons	Estimated proportion
Non-Affluent	500	145	0.29
Affluent	300	42	0.14.

use an appropriate hypothesis test to wheather there is any difference in proportion of customer who use discount coupons at $\alpha = 0.05$.

Ans:- The Null & Alternate hypothesis

$H_0 : P_1 = P_2$

$H_1 : P_1 \neq P_2$

$\bar{P}_1 = 0.29, \quad \bar{P}_2 = 0.14.$

pooled proportion $\bar{P} = \dfrac{n_1\bar{P}_1 + n_2\bar{P}_2}{n_1 + n_2}$

$= \dfrac{500 \times 0.29 + 300 \times 0.14}{500 + 300}$

$= \dfrac{145 + 42}{800} = \dfrac{187}{800}$

$= 0.2338.$

$$Z = \frac{(\bar{P}_1 - \bar{P}_2) - 0}{\sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.29 - 0.14}{0.2338[1 - 0.233}$$

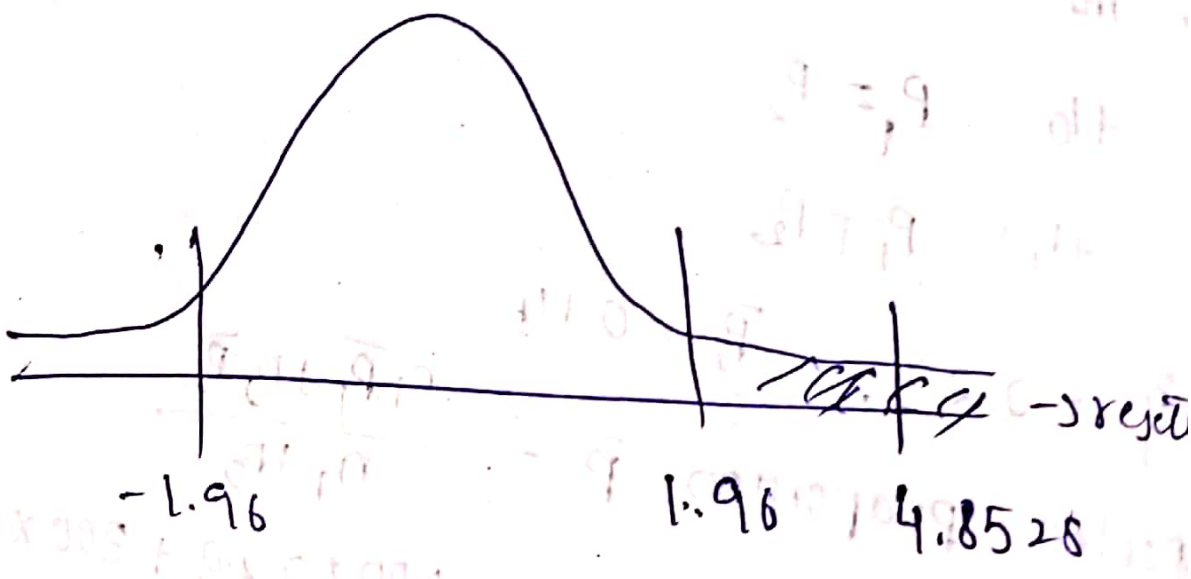
$$= \frac{0.29 - 0.14}{0.2338\left(1 - 0.2338\right)\left(\frac{1}{500} + \frac{1}{300}\right)}$$

$$= \frac{0.5}{0.179\left(0.002 + 0.0033\right)}$$

$$= \frac{0.15}{0.0391} = 4.8528$$

at 0.05 TWO tail z critical valu 1.96

So we can rejecte Null hypothesis



23/09/2019

F- Distribution : (Fisher's distribution)

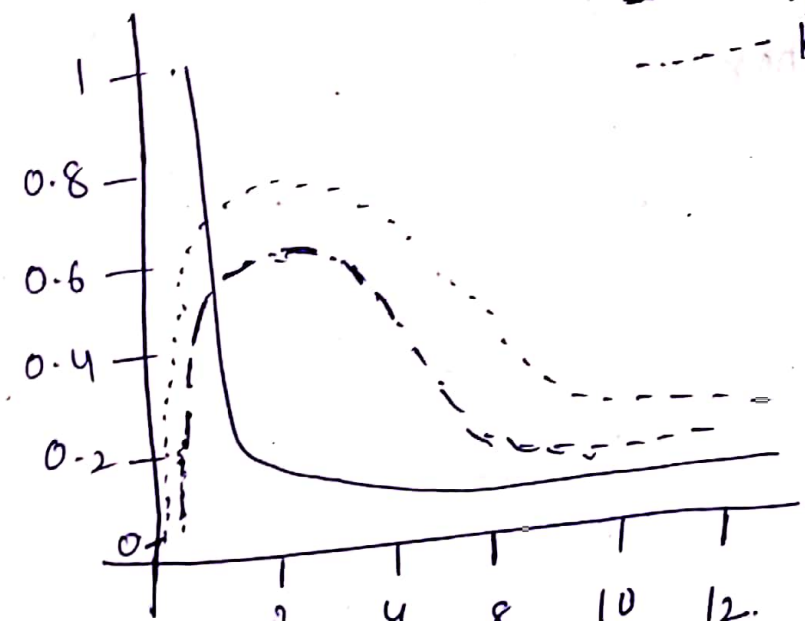
⇒ Ratio of two χ^2 distribution is follow F-Distribution.

⇒ F-Distribution is a ratio of two χ^2 distributions.

Let Q_1 and Q_2 be two Independent χ^2 distributions with k_1 & k_2 degree of freedom Respectively. then Random variable 'F' defined as $F = \frac{Q_1/k_1}{Q_2/k_2}$ is an 'F' Distribution.

Probability Density function

$$f(x) = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right)\left(k_1/k_2\right)^{k_1/2}}{\Gamma\left(\frac{k_1}{2}\right)+\left(\frac{k_2}{2}\right)} \times \frac{x^{\frac{k}{2}-1}}{\left(1+\frac{k_1-x}{k_2}\right)^{\frac{k_1+k_2}{2}}}$$



Properties 'f' Distribution:

1) Mean of 'f' distribution is $\frac{k_2}{(k_2-2)}$ $k_2 > 2$

2) Standard deviation of 'f' distribution is $\sqrt{\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}}$ for $k_2 > 4$

k_1, k_2 are degree of freedom

3) 'F' distribution is non-symmetrical and the shape of the distribution depends on values k_1 and k_2

4) 'F' distribution is used in Analysis of variance to Test the Mean values of Multiple groups

5) * With Two-Tail F Test we Just want to If the variance are not equal to each other

① conduct a Two-Tail 'F' Test and the following samples.

Sample 1

Variance = 109.63

$n_1 = 41$

Sample 2

Variance = 65.99

$n_2 = 21$.

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$

Calculate 'F' statistic

$$F = \frac{\sigma_1^2}{\sigma_1^2} \quad \frac{\Sigma(x - \bar{x}_1)^2}{(n-1)}$$

It will follow 'F' distribution.

$$F = \frac{109.63}{65.99} = 1.66.$$

Assume $\alpha = 0.05$ and degree freedom $df_1 = 40$

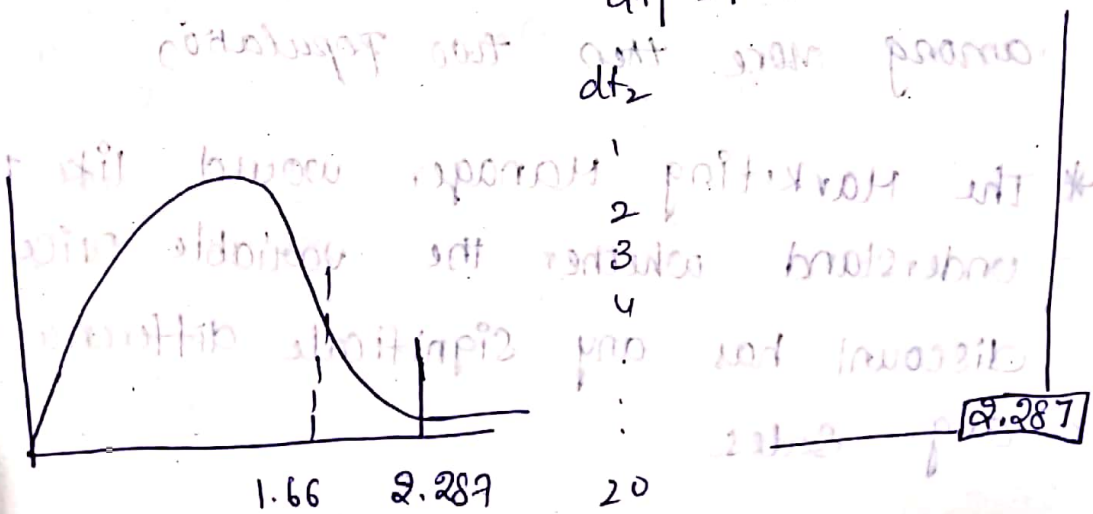
$df_2 = 20$

Table $\alpha = 0.025$

$df_1 = 1 \quad 2 \quad 3 \quad \cdots \quad \infty$

df_2

1
2
3
4
⋮
20



We can not reject null hypotheses since 'F' Statistic value less then critical value so we can not reject null hypothesis

Since two variance not significate differ

24/09/2019

Analysis of variance (ANOVA):-

⇒ In many situation we need to conduct a hypothesis to compare mean value simentainous in two groups (sample) created using a factor

Ex:- A Marketer like understand the impact of '3' different discount values such (0%, 10%, 20%) on the an average sales. when we have to compare the impact of a factor on mean on more then two groups simentainously the hypothesis test such two sample T-test not a Ideal approach since they can result in Incorrect Type I and Type II error. We can use the Analysis of variance to understand the difference in population means among more then two population

* The Marketing Manager would like to understand whether the variable price discount has any significate difference on avg sales.

* In ANOV our objective is to varity whether variance due to tretement is differene from variance due to randomness.

One way Analysis Variance:

→ one way ANOV Appropriate under following condition

* We would like to study Empact of single Treatment (factor) at differente Level on continouses Reponses variable (outcome).

For Example The variable Price discount is the tretement Factor and 0%, 10% and 20% price discounts are differente Levels (3 Levels) differente Levels of discounts are likely to have a varing Empact on Sales of the product where Sales is outcome variable.

* In Each Group the population response variable follows a normal distribution and the Sample Choosen using Random Sampling.

* Population variance for differente Groups are assume to be same. ie variability in the Responce variable values with in differente Groups is same

Setting up Analysis of Variance:

* Assume that we would like to study the Impact of a factor (discount) with k Levels on ~~Continuous~~ vo continuous variable (such as Sales Quntity). Then the null and Alternate hypothesis for one way ANOVA

H_0 : $\mu_0 = \mu_1 = \mu_2 = \cdots = \mu_k$

H_1 : Not all μ's are equal

① Using the following data perform a one way Analysis of ~~variable~~ variance using $\alpha = 0.05$

Group 1	Group 2	Group 3
51	23	56
45	43	76
33	23	74
45	43	87
67	45	56

$$\text{Group 1 Mean} = \frac{51+45+33+45+67}{5} = 48.2$$

$$\text{Group}_2 = \frac{23+43+23+43+45}{5} = 35.4$$

$$\text{Group}_3 = \frac{56+76+74+87+56}{5} = 69.8$$

GROUP, variance

	Mean	deviation	Squar deviat
51	48.2	51 − 48.2 = 2.8	$(2.8)^2 = 7.84$
45	48.2	−3.2	10.24
33	48.2	−15.2	231.04
45	48.2	−3.2	10.24
67	48.2	18.2	354.44

GROUP$_2$

		deviation	Squar.
23	35.4	−12.4	153.76
43	35.4	7.6	57.76
23	35.4	−12.4	153.76
43	35.4	7.6	57.76
45	35.4	9.6	92.16

GROUP$_3$

56	69.8	−13.8	190.44
76	69.8	6.2	38.44
74	69.8	4.2	17.64
87	69.8	17.2	295.84
56	69.8	−13.8	190.44

Sum of Square deviation (G_1) $= \dfrac{7.84 + 10.24 + 231.04 + 10.24 + 354.44}{5}$

$= 612.84$

$G_2 = \dfrac{153.76 + 57.76 + 153.76 + 57.76 + 92.16}{5}$

$= 515.2$

$G_3 = \dfrac{190.44 + 38.44 + 17.64 + 295.84 + 190.44}{5}$

$= 732.8$

$var_1 = 612.8 / 5-1 = \frac{612.8}{4} = 153.2$

$var_2 = \frac{515.2}{5-1} = \frac{515.2}{4} = 128.8$

$var_3 = \frac{732.8}{4} = 183.2$

$Mean\,Square_{Error}\ (MS_{Error}) = \frac{153.2 + 128.8 + 183.2}{3}$

$= 155.07.$

25/09/2019

(Grand μ) $(\mu_{Grand}) = \frac{48.2 + 35.4 + 69.8}{3} = 51.13$

Again find Variance for Group Means

Group Mean	Grand Mean	$\frac{\Sigma (x-\bar{x})^2}{n-1}$
48.2	51.13	=
35.4	51.13	
69.8	51.13	=

$\frac{\Sigma (x-\bar{x})^2}{n-1}$

$= \frac{(48.2-51.13)^2 + (35.4-51.13)^2 + (69.8-51.13)^2}{3-1}$

$= \frac{604.58}{2} = 302.29.$

(Mean Square Variation with in Group)] MSW

MSB (Mean Square Variation between groups)

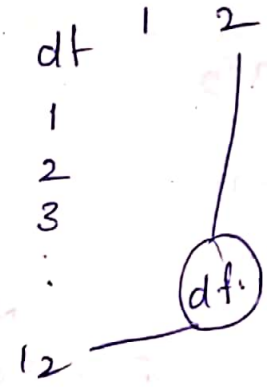
$= 302.29 \times 5$

$= 1511.45$

F- Statistic $= \frac{MSB}{MSW}$

$= \frac{1511.45}{155.07}$

$= 9.75$

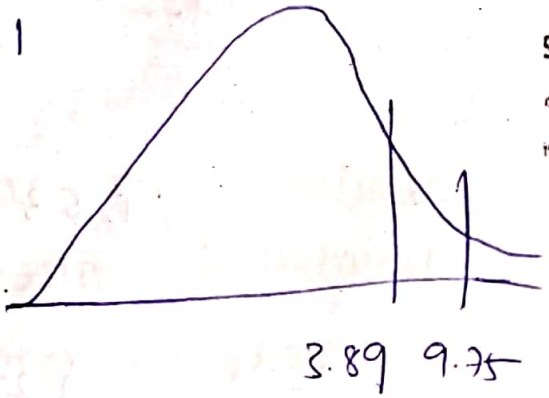


Degree of freedom within group $= 15 - 3$

$= 12$

Degree of freedom across all Group $= 3 - 1$

$= 2$



$df(2, 12) = 3.89$

reject null hypothesis

With in the following summary data Perfor
a one way Analysis of variance using

	n	mean	std. dev
G_1	30	50.26	10.45
G_2	30	45.32	12.76
G_3	30	53.67	11.47

Variance$_1$ $= 10.45^2 = 109.2$

$var_2 = 12.76^2 = 162.82$

$var_3 = 11.47^2 = 131.56$

$$F = \frac{MSB}{MSW}$$

$$MSW = \frac{109.2 + 162.82 + 131.56}{3}$$

$$= 134.53.$$

$$\text{Grand Mean} = \frac{50.26 + 45.32 + 53.67}{3}$$

$$= 49.75$$

Group mean

$$= \frac{(50.26 - 49.75)^2 + (45.32 - 49.75)^2 + (53.67 - 49.75)^2}{3 - 1}$$

$$= \frac{35.25}{2} = 17.62$$

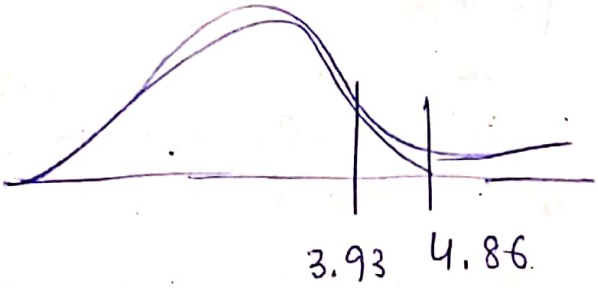
$\Rightarrow MSB = 17.62 \times 30$

$= 528.75$

$F\_statistic = \frac{MSW}{MSR}$

$= \frac{528.75}{134.53} = 3,93$

$F_{critica}(2, 87) = 4.86.$



All Group means are Equal. So we can not reject null hypotheses

A clenical Psychologes has run a between Subject Experiment comparing to tretments ~~has~~ for depression (-formative behavioral) and client center ther. against a control condition. Subjects for Randomly assign the expermental conditions after 12 weeks. The subjects differen Score Measure using CESD ~~scale~~ ditination Scale that data are summarized as.

	n	Mean	Sd.dev
Control	40	21.4	4.5
CBT	40	16.9	5.5
CCT	40	19.1	5.8

Using a one way Alova with $\alpha = 0.05$

$$MSW = \frac{20.25 + 30.25 + 33.64}{3} = 28.05$$

Grand Mean $= 19.13$.

$$\text{Group Mean} = \frac{(21.4 - 19.13)^2 + (16.9 - 19.13) + 19.1}{3}$$

$$= \frac{5.15 + 4.97 + 0.00}{2}$$

$$= \frac{10.12}{2}$$

$$MSB = \frac{10.12}{2} \times 40 = 202.4$$

$$F = \frac{MSW}{MSB} \qquad F = \frac{MSB}{MSW}$$

$$= \frac{28.05}{202.4} \qquad = \frac{202.4}{28.05}$$

$$= 7.22$$

$F_{Critical}(8; 117) = 4.79$

We can reject null hypothesis.

Covariance Analysis :-

26/09/19

Two variance indicate the direction of linear relationship between variables

Corelation :-

Corelation Measure both strength and direction of the linear relationship between two variables.

Covariance of two variables x, y. cov(x,y) can be represented as cov(x,y)

If E(x) is Expected value or Mean of a sample x. then

$$cov(x, y) = E[(x - \mu_x)(y - \mu_y)]$$

Corelation Co-efficent :- (Pearson corelation coefficent)

=> pearson product moment

is used for measuring the strength and direction of linear Relationship between two continuous random variables x, y.

Let x_i be different value of variable x and y_i be different value of variable y then pearson corelation coefficent

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_x \sigma_y}$$

When we try to Measure how change in variable [Y] is related to changing another variable (X) one of the Issues that we need Consider is the Measurement Scale and unit of Measurement of two variables

For Example Age is Measured in years. and call duration is Measured in Seconds. The Range of two variables can be different, Thus we need standaged variables which can be used for Measuring Corelation between two variables

$$z = \left(\frac{x_i - \bar{x}}{\sigma_x}\right) \qquad z_y = \left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

$$r = \sum_{i=1}^{n} \frac{z_x \cdot z_y}{n}$$

$$\gamma = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1} \sum \left(\frac{y_i - \bar{y})^2}{n-1}\right).}}$$

$$\gamma = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Properties:-

⇒ positive value indicate positive corelation.

⇒ Negative value indicate Negative Corelation

⇒ pearson corelation coefficient value may be 0 Even there is strong Non-relation ship between x, y Thus Low corelation coefficient value can not be Take An Eveidence no relationship.

① The Avg share raises of two companies over the first ~~6~~ 3 Months

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$
274.58	219.50	274.58 − 284.286	−16.6
287.96	242.92	−9.706	6.82
290.35	245.90	3.68	9.82
		−6.064	

$\bar{x} = 284.286$ $\bar{y} = 236.10$

$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) \times (y_i - \bar{y})$
94.08	275.56	161.02
13.54	46.51	25.09
36.84	96.43	59.4
144.42	418.1	245.6

$$r = \frac{245.6}{\sqrt{144.42} \times \sqrt{418.1}}$$

$$r = \frac{245.6}{12.01 \times 20.45}$$

$$r = \frac{245.6}{245.60}$$

$$\therefore r = 1$$

positively co-related and strongly Corelated

22/09/19

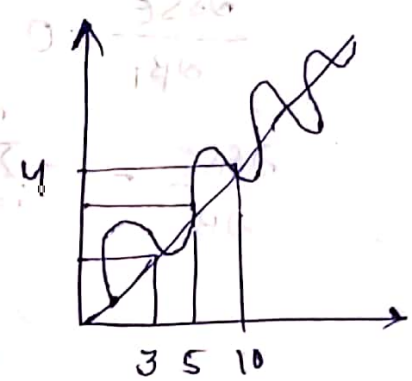
Simple Liner Regression:-

Simple linear regression is a statistical for finding the existance of an association blw the dependent variable and indepent variable

- Simple linear regression implies that there is only one independent variable in the order.

Functional form of SLR:-

$$y = \beta_0 + \beta_1 x + \varepsilon$$



* For a dataset with n observations (x_i, y_i) where $i = 1, 2, 3, \ldots n$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where ε - error

$\varepsilon = y_i - \beta_0 - \beta_1 \times$

* Estimation of parameters using ordinary least squares :-

$$\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - B_0 - B_1 x_i)$$

- In ordinary least squares the objective is find the optimal value of B_0 and B_1 that will minimize the sum of squares error (SSE)

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - B_0 - B_1 x_i)^2$$

- To find the optimal values of B_0 & B_1 that will minimize SSE, we have to equate the partial derivates of SSE with respect to B_0 & B_1 to 0.

$$\frac{\partial SSE}{\partial B_0} = 0$$

$$\frac{\partial SSE}{\partial B_0} = -\sum_{i=1}^{n} 2(y_i - B_0 - B_1 x_i) = 0$$

$$= 2(nB_0 + B_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i) = 0$$

$$\boxed{B_0 = \bar{y} - B_1 \bar{x}} \quad \text{---------- (1)}$$

$$\frac{\partial SSE}{\partial B_1} = 0$$

$$\frac{\partial SSE}{\partial B_1} = -\sum_{i=1}^{n} 2x_i (y_i - B_0 - B_1 x_i^2) = 0$$

$$= -2\sum_{i=1}^{n} (y_i x_i - B_0 x_i - B_1 x_i^2) = 0$$

Substitute B_0 value in above eqn

$$= -2\sum_{i=1}^{n} (y_i x_i - x_i \bar{y} + \beta_1 x_i \bar{x} - \beta_1 x_i^2)$$

$$= \sum_{i=1}^{n} (x_i y_i - x_i \bar{y}) - \beta_1 \sum_{i=1}^{n} (x_i^2 - x_i \bar{x}) = 0$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y})}{\sum_{i=1}^{n} (x_i^2 - x_i \bar{x})}$$

$$\beta_1 = \sum_{i=1}^{n} \frac{x_i (y_i - \bar{y})}{x_i (x_i - \bar{x})}$$

The above β_1 can be simplified as

Since $\sum_{i=1}^{n} (\bar{x}\bar{y} - y_i \bar{x}) = 0$ & $\sum_{i=1}^{n} (\bar{x}^2 - x_i \bar{x}) = 0$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y}) + 0}{\sum_{i=1}^{n} (x_i^2 - x_i \bar{x}) + 0}$$

$$= \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y}) + \sum_{i=1}^{n} (\bar{x}\bar{y} - y_i \bar{x})}{\sum_{i=1}^{n} (x_i^2 - x_i \bar{x}) + \sum_{i=1}^{n} (\bar{x}_i^2 - x_i \bar{x})}$$

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{cov}(x,y)}{\text{var}(x)}$$

23/10/2019

Validation of Simple Linear Regression:

1. Co-Efficient of Determinant (R^2)
2. Hypothesis test for regression coefficient B_1
3. ANOVA (MLR)
4. Residual Analysis
5. Outlier Analysis

Hypothesis testing for B_1:-

$$R^2 = \frac{\text{Explained variance}}{\text{Total variation}}$$

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i$$

$\downarrow$

Variation in y_i = variation explained by model

SST = Sum of Square of total variation

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SST = SSR + SSE (Sum of Squares of Errors)

$$R^2 = \frac{\text{Explained variance}}{\text{Total variation}} = \frac{SR}{SST} = \frac{SST - SSE}{SST}$$

$$R^2 = 1 - \frac{SSE}{SST}$$

* R^2 is the proportion of variation in explained by the