Unit-1

Correlation:- Relationship between two variables such that a change in one variable results in a positive (or) negative change in the other variable.

There are three types of Correlation:-

1. positive Correlation
2. negative Correlation
3. perfect correlation

1. positive Correlation :- If two variables deviates in the same direction the correlation is said to be positive (or) direct Correlation
   * One variable increase the other variable also increase ↑↑ (or) ↓↓
   Example:- pressure cooker.

2. Negative Correlation:- If the variables deviate in the opposite directions the correlation is said to be inverse (or) negative correlation.
   * If one variable increases the other decrease ↑↓(or)↓↑
   Example:- Boils law, Exercise & Calories.

3. Perfect correlation:- If the change in one variable corresponds to a proportion change in the other variable, then the correlation is said perfect.

Karl pearsons cofficient of correlation:-

The numerical measure of linear relationship between the random variables (rv) X, Y is called cofficient of correlation between the variables and we write as R_{xy} (or) R

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \quad \text{(covariance)}$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

where

$x = x - \bar{x}$, $y = y - \bar{y}$ (observations are x and y & mean $\bar{x}, \bar{y}$)

$$Cov(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\left[\frac{1}{n} \Sigma (x_i - \bar{x})^2 * \Sigma (y_i - \bar{y})^2\right]^{1/2}$$

Note:- 1. 'r' lies between −1 and +1

2. $r = 0$, variables are not correlated

(or)

Variables are Uncorrelated

3. $r = +1$, Variables positively perfectly correlated

4. $r = -1$, Variables negatively perfectly correlated.

Problem:- calculate the coefficient of correlation for the heights of fathers & their sons?

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	x^2	y^2	XY
65	67	65-68 = -3	67-69 = -2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8

$\Sigma x = 544$ $\Sigma y = 552$ $\Sigma x^2 = 36$ $\Sigma y^2 = 44$ $\Sigma XY = 24$

$\bar{x} = \frac{544}{8} = 68$ $\bar{y} = \frac{552}{8} = 69$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{24}{\sqrt{36 \times 44}}$$

$$= \frac{24}{6\sqrt{44}}$$

$$= \frac{4}{2\sqrt{11}}$$

$$= \frac{2}{\sqrt{11}}$$

$$r = 0.6030$$

Problem:- Marks obtained by the students in maths and statistics by 10 students is given below. Find the correlation between two subjects

Marks in Maths	75	30	60	80	53	35	15	40	38	48
Marks in statistics	85	45	54	91	58	63	35	43	45	44

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
75	85	$75 - 47.4 = 27.6$	$85 - 56.3 = 28.7$	761.76	823.69	792.12
30	45	-17.4	-11.3	302.76	127.69	196.62
60	54	12.6	-2.3	158.76	5.29	-28.98
80	91	32.6	34.7	1062.76	1204.09	1131.22
53	58	5.6	1.7	31.36	2.89	9.52
35	63	-12.4	6.7	153.76	44.89	-83.08
15	35	-32.4	-21.3	1049.76	453.69	690.12
40	43	-7.4	-13.3	54.76	176.89	98.42
38	45	-9.4	-11.3	88.36	127.69	106.22
48	44	0.6	-12.3	0.36	151.29	-7.38
$\Sigma x = 474$	$\Sigma y = 563$			$\Sigma x^2 = 3664.4$	$\Sigma y^2 = 3118.3$	$\Sigma XY = 2904.8$

$\bar{x} = \frac{474}{10} = 47.4$

$\bar{y} = \frac{563}{10} = 56.3$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{2904.8}{\sqrt{(3664.4) * (3118.3)}}$$

$$= \frac{2904.8}{3380.34}$$

$$r = 0.8593$$

* 5th decimal place ≥ 5 add 1 to previous value
* 5th decimal place < 5 the previous number is same

correlation coefficient - another method

Direct method:-

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} * \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

where n = no of observations

Problem:- Find 'r' for the following data by direct method.

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

x	y	xy	x^2	y^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
$\Sigma x = 544$	$\Sigma y = 552$	$\Sigma xy = 37560$	$\Sigma x^2 = 37028$	$\Sigma y^2 = 38132$

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} * \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$r = \frac{8*(37560) - (544)(552)}{\sqrt{8*(37028) - (544)^2} * \sqrt{8(38262) - (552)^2}}$$

$$= \frac{300480 - 300288}{\sqrt{296224 - 295936} * \sqrt{306096 - 304704}}$$

$$= \frac{192}{\sqrt{288} * \sqrt{1392}}$$

$$= \frac{192}{\sqrt{400896}}$$

$$\neq \frac{192}{633.1635} = 318/318.3960$$

$r = 0.6030$

② Problem:- Find the correlation cofficient for the following data by direct method

X	20	30	40	80	60	70
y	23	18	30	20	15	8

x	y	xy	x^2	y^2
20	23	460	400	529
30	18	540	900	324
40	30	1200	1600	900
80	20	1600	6400	400
60	15	900	3600	225
70	8	560	4900	64
$\Sigma x = 300$	$\Sigma y = 114$	$\Sigma xy = 5260$	$\Sigma x^2 = 17800$	$\Sigma y^2 = 2442$

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} * \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{6*(5260) - (300)(114)}{\sqrt{6*57800 - (300)^2}\ \sqrt{6*2442 - (114)^2}}$$

$$= \frac{31560 - 34200}{\sqrt{106800 - 90000} * \sqrt{14652 - 12996}}$$

$$= \frac{-2640}{\sqrt{16800}\ \sqrt{1656}}$$

$$= \frac{-264\cancel{0}}{10\sqrt{168} * \sqrt{1656}}$$

$$= \frac{-264}{\sqrt{278208}}$$

$$\boxed{r = 0.5005}$$

Regression:- Regression means stepping back towards the normal

There are two types of Regression lines

1. Regression line Y on X
2. Regression line X on Y

1. Regression line Y on X:-

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

where $r \frac{\sigma_y}{\sigma_x} = b_{yx}$ is the regression cofficient of Y on X

σ_x = standard deviation of X

σ_y = standard deviation of Y

$\bar{x}$ = mean of X

$\bar{y}$ = mean of Y

2. Regression line X on Y:-

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

where $r \frac{\sigma_x}{\sigma_y} = b_{xy}$ is the regression cofficient of X on Y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

σ_x = SD of x

σ_y = SD of y

Note:- $r^2 = b_{yx} * b_{xy}$ (01)

$r = \sqrt{b_{yx} * b_{xy}}$

Problem:- Calculate the cofficient of correlation from the following data. Also obtain the equation of the lines of regression & obtain an estimate of y which should correspond on the average of $x = 6.2$

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Here $n = 9$

x	y	$x = x - \bar{x}$	$y = y - \bar{y}$	x^2	y^2	xy
1	9	-4	-3	16	9	12
2	8	-3	-4	9	16	12
3	10	-2	-2	4	4	4
4	12	-1	0	1	0	0
5	11	0	-1	0	1	0
6	13	1	1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	12
$\Sigma x = 45$	$\Sigma y =$			$\Sigma x^2 = 60$	$\Sigma y^2 = 60$	$\Sigma xy = 57$
$= 5$	$108 = 12$					

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{57}{\sqrt{60 * 60}} = \frac{57}{60} = 0.95$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{n}\Sigma x^2}$$

$$= \sqrt{\frac{1}{n} * 60}$$

6.6667

$$= \sqrt{\frac{1}{9} * 60}$$

$$= \sqrt{\frac{20}{3}}$$

$$\boxed{\sigma_x = 2.5819}$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma(y_i - \bar{y})^2}$$

$$= \sqrt{\frac{1}{9}\Sigma y^2}$$

$$= \frac{1}{3}\sqrt{60}$$

$$= \frac{1}{3} * 7.7460$$

$$\boxed{\sigma_y = 2.5819}$$

Regression line Y on X:-

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 12 = 0.95 \left(\frac{2.5819}{2.5819}\right)(x - 5)$$

$$y - 12 = 0.95 (x - 5)$$

$$y - 12 = 0.95x - 4.75$$

$$\boxed{y = 0.95x + 7.25}$$

Regression line X on Y:-

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 5 = 0.95 * \left(\frac{2.5819}{2.5819}\right)(y - 12)$$

$$x - 5 = 0.95 (y - 12)$$

$$x - 5 = 0.95y - 11.4$$

$$\boxed{x = 0.95y - 6.4}$$

Problem:- The following data relate to the marks of 10 students in the internal test and the university examination for the maximum of 50 each.

Internal Exam	25	28	30	32	35	36	38	39	42	45
University marks	20	26	29	30	25	18	26	35	35	46

1. The most likely internal mark for univers
   mark of 25
2. The most likely university mark for interna
   mark of 30.

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
25	20	−10	−9	100	81	90
28	26	−7	−3	49	9	21
30	29	−5	0	25	0	0
32	30	−3	1	9	1	−3
35	25	0	−4	0	16	0
36	18	1	−11	1	121	−11
38	26	3	−3	9	9	−9
39	35	4	6	16	36	24
42	35	7	6	49	36	42
45	46	10	17	100	289	170
$\Sigma x = 350$	$\Sigma y = 290$			$\Sigma X^2 =$	$\Sigma Y^2 =$	$\Sigma XY =$
$\frac{350}{10} = 35$	$\frac{290}{10} = 29$			358	598	274

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{324}{\sqrt{358 * 598}}$$

$$= \frac{324}{\sqrt{214084}}$$

$$r = 0.7002$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{n}} \qquad \sigma_y = \sqrt{\frac{\Sigma y^2}{n}}$$

$$= \sqrt{\frac{358}{10}} \qquad = \sqrt{\frac{598}{10}}$$

$$= \sqrt{35.8} \qquad \sigma_y = 7.7330$$

$$\sigma_x = 5.9833$$

Regression y on x:-

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 29 = 0.7002 \left(\frac{7.7330}{5.9833} \right) (x - 35)$$

$$y - 29 = 0.7002 (1.2924) (x - 35)$$

$$y - 29 = 0.9050 (x - 35)$$

$$y - 29 = 0.9050x - 31.6736$$

$$y = 0.9050x - 2.6736$$

Regression x on y:-

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 35 = 0.7002 * \left(\frac{5.9833}{7.7330} \right) (y - 29)$$

$$x - 35 = 0.7002 * (0.7737) (y - 29)$$

$$x - 35 = 0.5418 (y - 29)$$

$$x - 35 = 0.5418y - 15.7122$$

$$x = 0.5418y + 19.2878$$

Estimation:-

$x \rightarrow$ internal mark

y - university mark

1. x when $y=25$ use x on y Regression Line

$$x = 0.5418(25) + 19.2878$$

$$x = 13.545 + 19.2878$$

$$\boxed{x = 32.8328}$$

2. y when $x=30$

use y on x Regression line

$$y = 0.9050(30) - 2.6736$$

$$= 27.15 - 2.6736$$

$$\boxed{y = 24.4764}$$

Problem:- 10 observations on price (x) and supply (y) data were obtained below

$\Sigma x = 130$, $\Sigma x^2 = 2288$, $\Sigma xy = 3467$

$\Sigma y = 220$, $\Sigma y^2 = 5506$

obtained the line of regression of y on x & Estimate the supply when the price is 16 units & also find the standard Error of the Estimate.

$$\boxed{r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}}}$$

Here, $n=10$

$$r = \frac{n*(\Sigma XY) - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$r = \frac{10(3467) - (130)(220)}{\sqrt{10(2288) - (130)^2}\ \sqrt{10(5506) - (220)^2}}$$

$$= \frac{34670 - 28600}{\sqrt{22880 - 16900}\ \sqrt{55060 - 48400}}$$

$$= \frac{6070}{\sqrt{5980}\ \sqrt{6660}}$$

$$= \frac{6070}{\sqrt{39826800}}$$

$$= \frac{6070}{6310.8478}$$

$$\boxed{r = 0.9618}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{130}{10} = 13, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{220}{10} = 13$$

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

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{n}\left[\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2\right.}$$

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

$$= \frac{1}{n}\left[\Sigma(x)^2 + \Sigma(\bar{x})^2 - 2\Sigma x\bar{x}\right]$$

$$= \frac{1}{n}\left[\Sigma x^2 + n(\bar{x})^2 - 2(x)n(\bar{x})\right]$$

$$= \frac{1}{n}\left[\Sigma x^2 - 2n\bar{x}^2 + n(\bar{x})^2\right]$$

$$= \frac{1}{n}\left[\Sigma x^2 - n(\bar{x})^2\right]$$

$$= \frac{1}{n}\left[\Sigma x^2 - n\left(\frac{\Sigma x}{n}\right)^2\right]$$

$$\sigma_x = \frac{1}{n}\left[\Sigma x^2 - (\Sigma x)^2\right]$$

$$\sigma_x = \sqrt{\frac{1}{10}\left[2288 - \left[\frac{136}{10}\right]^2\right]}$$

$$= \sqrt{\frac{1}{10}\left[2288 - \left[\frac{1690}{10}\right]\right.}$$

$$= \sqrt{\frac{1}{10}\left[\left[2288\right] - 1690\right.}$$

$$= \sqrt{\frac{211}{10}} \quad \sqrt{59.8}$$

$\sigma_x = \sqrt{211.9}$

σ_x

$\sigma_x = 7.7330$

$$\sigma_y = \sqrt{\frac{1}{n}\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}$$

$$= \sqrt{\frac{1}{n}\left[5506 - \frac{(220)^2}{10}\right]}$$

$$= \sqrt{\frac{1}{10}\left[5506 - \frac{48400}{10}\right]}$$

$$= \sqrt{\frac{1}{10}\left[5506 - 4840\right]}$$

$$= \sqrt{666/10} \qquad \boxed{\sigma_y = 8.1608}$$

Estimation:-

y when $x = 16$

$y = 1.0150 * 16 + 8.8049$

$y = 16.24 + 8.8049$

$y = 25.0449$

Standard Error:-

$$S.E = \frac{1 - r^2}{\sqrt{n}}$$

$$S.E = \frac{1 - (0.9618)^2}{\sqrt{10}}$$

$$= \frac{1 - 0.92505}{\sqrt{10}}$$

$$= \frac{1 - 0.9250}{\sqrt{10}}$$

$$= \frac{1 - 0.2925}{}$$

$$S.E = 0.0237$$

***

Problem:- The eqn of two regression lines obtained in a correlation analysis

$3x + 12y = 19$, $3y + 9x = 46$

a) Find cofficient of correlation

b) Mean values of x & y

c) Ratio of coffecients of variability.

Sol^{n}:- Let

$3x+12y=19$ ——— (1)

$3y+9x=46$ ——— (2)

nothing is specified as y on x (or) x on y than consider (1) eq^{n} y on x and (2) eq^{n} x on y

i.e; $3x+12y=19$

$12y=-3x+19$

$y=\frac{-3x}{12}+\frac{19}{12}$ $(\because y=mx+c)$

$y=\frac{-x}{4}+\frac{19}{12}$ eq^{n} of line

m ⟶ slope is an regression coefficient

$b_{yx}=\frac{-3}{12}=-1/4$

Let suppose eq^{n} (2) to be x on y regression Line

i.e; $3y+9x=46$

$9x=-3y+46$

$x=\frac{-3y}{9}+46$

$x=\frac{-y}{3}+\frac{46}{3}$

$b_{xy}=-1/3$ is the regression coefficient

$r=\sqrt{b_{yx}*b_{xy}}$

$=\pm\sqrt{(-1/4)(-1/3)}$

$r=-0.2886$

when Ever byx and bxy are '-'ve & 'r' is '-'ve.
If byx & bxy are '+'ve than 'r' is '+'ve
Range of 'r' is '-1 to +ive.

b) Mean value:-

Mean value of x & y

solving eqn ① & ② we get

$3x + 12y = 19$ —— ①

$9x + 3y = 48$ —— ②

$9x + 36y = 57$

$9x + 3y = 46$

$33y = 11$

$y = 11/33$

$y = 1/3$

$3x + 12y = 19$

$36x + 12y = 184$

$-33x = -165$

$x = \frac{165}{33}$

$x = 5$

Intersection of point of two regression lines is the mean value of x & y.

$(5, 1/3) = (\bar{x}, \bar{y}) = (5, 0.3333)$

3. Coefficient of variability:-

$CV = \frac{\sigma x}{\sigma y}^2$

we have $bxy = r\frac{\sigma x}{\sigma y}$

$(bxy)^2 = \frac{r^2 \sigma x^2}{\sigma y^2}$

$\frac{\sigma x^2}{\sigma y^2} = \left(\frac{bxy}{r}\right)^2 = \frac{(-1/3)^2}{1/12} = \frac{1}{9/3} \times 1/1^4$

$r = 4/3$

$r = 1.3333$

Coff of variability $= 1.3333$

19/10/2020
Monday

** Rank correlation:- Suppose $x_1, x_2, \ldots x_n$ and $y_1, y_2, \ldots y_n$ be the ranks of two variables x & y in the order of merit with respect to some property. The correlation b/w these in pairs of ranks is called rank correlation.

Rank correlation is measured by spearman's rank correlation cofficient.

$$\rho = 1 - \left[\frac{6 * \left(\sum_{i=1}^{n} di^2 \right)}{n(n^2-1)} \right]$$

where n = no. of observation
x_i = Rank of X
y_i = Rank of Y
$di = x_i - y_i$

Note:- The rank correlation cofficient (ρ) lies b/w -1 & 1

Problem:- obtain the Rank Correlation Coefficient for the following data

X	10	15	12	17	13	16	24	14	22	20
Y	30	42	45	46	33	34	40	35	39	38

X	Y	Rank of X $= x_i$	Rank of Y $= y_i$	$d_i = x_i - y_i$	d_i^2
10	30	10	10	0	0
15	42	6	3	3	9
12	45	9	2	7	49
17	46	4	1	3	9
13	33	8	9	-1	1
16	34	5	8	-3	9
24	40	1	4	-3	9
14	35	7	7	0	0
22	39	2	5	-3	9
20	38	3	6	-3	9
$\Sigma X = 163$	$\Sigma Y = 382$	$\Sigma x_i = 55$	$\Sigma y_i = 55$		$\Sigma d_i^2 = 104$

$$\rho = 1 - \left[\frac{6 * \Sigma d_i^2}{n}\right]$$

$$\rho = 1 - \left[\frac{6 * (\Sigma d_i^2)}{n(n^2-1)}\right]$$

$$= 1 - \left[\frac{6 * 104}{10(10^2-1)}\right]$$

$$= 1 - \left[\frac{6 * 104}{10(99)}\right]$$

$$P = 1 - \frac{624}{990}$$

$$= 1 - 6303$$

$$\boxed{P = 0.3697}$$

Holds good for Unique ranks.

Problem:- Calculate the rank correlation coefficient for the following data

X 124 100 105 112 102 93 99 115 123 104 98 113 121 103 101

Y 80 100 102 91 92 111 109 98 89 104 113 105 97 106 123

X	Y	Rank of X $= x_i$	Rank of Y $= y_i$	$x_i - y_i$	d_i^2
124	80	1	15	-14	196
100	100	12	9	3	9
105	102	7	8	-1	1
112	91	6	13	-7	49
102	92	10	12	-2	4
93	111	15	3	12	144
99	109	13	4	9	81
115	98	4	10	-6	36
123	89	2	14	-12	144
104	104	8	7	1	1
98	113	14	2	12	144
113	105	5	6	-1	1
121	97	3	11	-8	64
103	106	9	5	4	16
101	123	11	1	10	100

$\Sigma di^2 = 990$

$$\rho = 1 - \left[\frac{6 * \Sigma di^2}{n(n^2-1)}\right]$$

$$= 1 - \left[\frac{6 * 990}{15(15^2-1)}\right]$$

$$= 1 - \left[\frac{6 * 990}{15(224)}\right]$$

$$= 1 - \left[\frac{5940}{3360}\right]$$

$$= 1 - [1.7679]$$

$$\boxed{\rho = 0.7679}$$

Repetation of ranks :-

$$\boxed{\rho = 1 - \left[\frac{6 * [\Sigma di^2 + C.F]}{n(n^2-1)}\right]}$$

CF = Correlation Factor

CF = CF in X series + CF in Y series

Formula for

$$\boxed{CF = \frac{m(m^2-1)}{12}}$$

where m is the no. of times the rank is repeated

Problem:- Calculate the rank correlation coefficient for the following data

(24)

X	Y	x_i	y_i	$x_i - y_i$	d_i^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16

$\Sigma d_i^2 = 72$

$$\rho = 1 - \left[\frac{6 * \left[\Sigma d_i^2 + CF \right]}{n(n^2-1)} \right]$$

CF = CF in X series + CF in Y series

CF in X series:- 75 is repeated 2 times
64 is repeated 3 times

$$CF = \frac{(2^2-1)}{12} + \frac{(3^2-1)}{12}$$

$$= \frac{4-1}{12} + \frac{9-1}{12}$$

$$= \frac{3}{12} + \frac{8}{12}$$

$$CF = 1/2 + 2$$

$$= 2.5$$

CF in Y series:-

$$CF \text{ in } Y \text{ series} = \frac{2(2^2-1)}{2}$$

$$= 0.5$$

$$CF = CF \text{ in } X \text{ series} + CF \text{ in } Y \text{ series}$$

$$= 2.5 + 0.5$$

$$CF = 3.0$$

$$\rho = 1 - \left[\frac{6 * (72+3)}{10(10^2-1)}\right]$$

$$= 1 - \left[\frac{6 * (75)}{10 \times 99}\right]$$

$$= 1 - \left[\frac{450}{10 \times 99}\right]$$

$$= 1 - \left[\frac{45}{99}\right]$$

$$= 1 - 0.4545$$

$$\boxed{\rho = 0.5455}$$

Problem:- A sample of 12 fathers and their eldest sons gave the following data about their heights. Calculate their rank correlation.

X	Y	x_i	y_i	$x_i - y_i$	d_i^2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1	1
64	65	10	12.5	-1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	-3.5	12.25
66	65	8	11.5	-3.5	12.25
68	71	4.5	1	3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	-2.5	6.25
71	70	1	2	-1	1
					$\sum d_i^2 = 72.5$

CF in X Series:- 67 - 2 times
68 - 2 times

$$CF_X = \left(\frac{2(2^2-1)}{12}\right) + 2\left(\frac{2^2-1}{12}\right)$$

$$= 0.5 + 0.5$$

$$= 1$$

CF in Y series:- 68 - 4 times
65 - 2 times
66 - 2 times

$$CF_Y = \frac{4(4^2-1)}{12 \times 3} + \frac{2(2^2-1)}{12 \times 6} + \frac{2(2^2-1)}{12 \times 6}$$

$$= \frac{15}{8} + \frac{2}{12} + \frac{2}{12}$$

$CF_y = 5 + 0.5 + 0.5$

$CF_y = 6$

$CF = CF_x + CF_y$

$= 1 + 6$

$\boxed{CF = 7}$

$$\rho = 1 - \left[\frac{6(72.5+7)}{12(144-1)}\right]$$

$$= 1 - \left[\frac{\cancel{6}^{1}(\cancel{43.25}^{79.5})}{\cancel{12}^{2}(143)}\right]$$

$$= 1 - \left[\frac{\cancel{43.25}^{79.5}}{286}\right]$$

~~$= 1 - [0.1512]$~~

~~$\rho = 1 - 0.2789$~~

$$\rho = \left[1 - \frac{79.5}{286}\right]$$

$= 1 - 0.27797$

$\boxed{\rho = 0.7221}$

Limits of Correlation:-

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

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

Let $a_i = x_i - \bar{x}$

$b_i = y_i - \bar{y}$

$$r^2_{x,y} = \left(\frac{1}{n}\right)^2 \left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2$$

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

$$= \frac{\frac{1}{n^2}\left(\sum_{i=1}^{n} a_i b_i\right)^2}{\frac{1}{n^2}\left(\sum_{i=1}^{n} a_i^2 \sum b_j^2\right)}$$

$$r^2_{x,y} = \frac{\sum_{i=1}^{n}(a_i b_i)^2}{\sum a_i^2 \sum b_i^2} \rightarrow (1)$$

we have Cauchy Scwartz inequality which states that if $a_i b_i, i = 1,2,\ldots n$ are real quantities then

$$\left(\sum a_i b_i\right)^2 \leq \left(\sum a_i\right)^2 \left(\sum b_i^2\right)$$

The sign of equality holding if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$$

Using Cauchy Schwartz inequality in (1) we have $r^2(x,y) \leq 1$

$$\Rightarrow |r(x,y)| \leq 1$$

$$\boxed{-1 \leq r \leq 1}$$

Properties of correlation Cofficient:-

1. Limits for correlation Cofficient are (-1, +1)
2. Correlation Cofficient is independent of change of Origin & Scale.
3. Two independent Variables are Uncorrelated

Properties of Regression Cofficient:-

1. Correlation cofficient is the geometric mean between the regression cofficients

Proof:- we have $b_{yx} * b_{xy} = \frac{r\sigma_y}{\sigma_x} * \frac{r\sigma_x}{\sigma_y}$

$= r * r$

$b_{yx} * b_{xy} = r^2 \quad = r^2$

$$\boxed{r = \pm\sqrt{b_{yx} * b_{xy}}}$$

Note:- Sign of 'r' is same as that of regression cofficients

(∵ the sign of Each regression cofficient depends on the covariance term)

If both the regression coefficients are +ve then 'r' is +ve

If both the regression coefficients are -ve then 'r' is -ve.

2. If one of the regression coefficients is greater than Unity the other is less than Unity

Proof:- Let b_{yx} is greater than Unity we have to show that b_{xy} is less than Unity

$$b_{xy} > 1$$

$$\frac{1}{b_{xy}} < 1 \quad \text{---} (1)$$

Also $r^2 \leq 1 \Rightarrow b_{yx} b_{xy} \leq 1$

$$\Rightarrow b_{xy} \leq \frac{1}{b_{yx}}$$

$$\Rightarrow b_{xy} < 1 \quad (\text{from eqn } (1))$$

$\therefore$ If $b_{yx} > 1$ then $b_{xy} < 1$ & Vice versa.

3. The modulus of the arithematic mean of the regression coefficient is not less than the modulus value of the correlation coefficient 'r'.

i.e; $\left| \frac{b_{yx} + b_{xy}}{2} \right| > |r|$

Proof:- To prove that

$$\left| \frac{1}{2} (b_{yx} + b_{xy}) \right| > |r|$$

$$\Rightarrow \left| \frac{1}{2}\left(\frac{r\sigma_y}{\sigma_x}\right) + \left(\frac{r\sigma_x}{\sigma_y}\right) \right| > |r|$$

$$\Rightarrow \left| r\left(\frac{\sigma_y}{\sigma_x}\right) + r\left(\frac{\sigma_x}{\sigma_y}\right) \right| > 2|r|$$

$$\Rightarrow \left| \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \right| > 2 \quad (\because |r| > 0)$$

$$\Rightarrow \frac{\sigma_y^2 + \sigma_x^2}{\sigma_x \sigma_y} > 2 \quad (\because \sigma_x + \sigma_y > 0 \text{ modulus removed})$$

$$\Rightarrow \sigma_y^2 + \sigma_x^2 > 2\sigma_x\sigma_y$$

$$\Rightarrow \sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y > 0$$

$$\Rightarrow (\sigma_y - \sigma_x)^2 > 0$$

which is always true since the square of a real quantity is >0.

4. The regression coefficients are independent of change of origin but not scale.

Proof:- let $u = \frac{x-a}{h}$ & $v = \frac{y-b}{k}$ where a, b, h & k are constants.

$$\Rightarrow x = a + uh \; ; \; y = b + vk$$

$$Cov(x,y) = E\left[\{x - E(x)\}\{y - E(y)\}\right]$$

$$= E\left[\{(a+uh) - E(a+uh)\}\right] * \left[(b+vk) - E(b+vk)\right].$$

$$Cov(X,Y) = E\left[h\{U - E(U)\}\{k[V - E(V)]\}\right]$$

$$= hk\,E\left[\{U - E(U)\}\{V - E(V)\}\right]$$

$$Cov(X,Y) = hk\,cov(U,V) \quad \text{---} (1)$$

$$Var(X) = \sigma_x^2 = V(X) = Var(a+uh)$$

$$= h^2 Var(U) \rightarrow (2)$$

Similarly

$$\sigma_y^2 = V(Y) = Var(b+kv)$$

$$= k^2 Var(V)$$

we have $b_{yx} = \dfrac{cov(X,Y)}{\sigma_x^2}$

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{yx} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} * \frac{\sigma_y}{\sigma_x}$$

$$\boxed{b_{yx} = \frac{cov(X,Y)}{\sigma_x^2}}$$

$$b_{yx} = \frac{cov(X,Y)}{\sigma_x^2} = \frac{hk\,cov(U,V)}{h^2(V(U))} \quad (\text{from } (1) \& (2))$$

$$= \frac{k}{h}\,\frac{co(U,V)}{Var(U)}$$

$$\boxed{b_{yx} = \frac{k}{h}\, b_{vu}}$$

Similarly we can prove that $b_{xy} = \frac{h}{k} b_{uv}$

∴ Regression coefficient are independent of change of Origin but not scale

Angle between two regression Lines:-

Let $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \longrightarrow (1)$ be the regression line y on x

Let $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \longrightarrow (2)$ be the regression line x on y.

$(2) \Rightarrow y - \bar{y} = \frac{\sigma_y}{r\sigma_x}(x - \bar{x})$

Angle between the lines of regression:-

$$\tan\theta = \left| \frac{r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + \left(r\frac{\sigma_y}{\sigma_x}\right)\left(\frac{\sigma_y}{r\sigma_x}\right)} \right|$$

$$= \left| \frac{r^2\sigma_y - \sigma_y}{r\sigma_x} * \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} \right|$$

$$= \left| \frac{\sigma_y(r^2 - 1)}{r\sigma_x} * \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} \right|$$

$$\tan\theta = \frac{\sigma_x \sigma_y (r^2 - 1)}{r(\sigma_x^2 + \sigma_y^2)}$$

$$\tan\theta = \left|\frac{(r^2-1)}{r}\right| \left|\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right|$$

$$\tan\theta = \left(\frac{|1-r^2|}{|r|}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right) \quad (\because r^2 \leq 1)$$

$$\boxed{\theta = \tan^{-1}\left\{\left(\frac{1-r^2}{|r|}\right)\left(\frac{\sigma_x \sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\}}$$

Case 1:- If $r=0$, $\tan\theta = \infty \Rightarrow \theta = \pi/2$

Thus if the two variables are uncorrelated the lines of regression becomes $\perp$ar to each other.

Case 2:- If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0 (or) \pi$

In this case, the two lines of regression either coincide (or) they are parallel to each other. since lines of regression pass through $(\bar{x}, \bar{y})$, they can't be parallel. so, in case of perfect correlation +'ve(or) -'ve, the two lines of regression coincide.

Problem:- The data given below relates to marks in two subjects mathematics and statistics of B.Tech students. The correlation coefficient b/w the marks in two subjects is 0.42

1. Estimate the marks in statistics if the marks in maths is 52
2. Find the angle b/w two regression lines

	Maths x	statistics y
Average marks means	39.5	49.5
standard deviation	10.8	16.8

The Regression line Y on X:-

$$y-\bar{y} = r\frac{\sigma y}{\sigma x}(x-\bar{x})$$

$$y-49.5 = (0.42)\left(\frac{16.8}{10.8}\right)(x-39.5)$$

$$y-49.5 = (0.42)(1.5556)(x-39.5)$$

$$y-49.5 = 0.6533(x-39.5)$$

$$y-49.5 = 0.6533x - 25.\frac{8074}{8057}$$

$$y = 0.6533x - 25.\frac{8057}{8057} + 49.5$$

$$y = 0.6533x + 23.6933$$

when $x = 52$

$$y = 0.6533 * 52 + 23.6933$$

$$y = 33.9716 + 23.6933$$

$$y = 57.6649$$

$$\theta = \tan^{-1}\left\{\left(\frac{1-r^2}{1r}\right)\left(\frac{\sigma x\, \sigma y}{\sigma x^2 + \sigma y^2}\right)\right\}$$

$$= \tan^{-1}\left\{\left(\frac{1-0.42)^2}{0.42}\right)\left(\frac{10.8 * 16.8}{(10.8)^2 + (16.8)^2}\right)\right.$$

$$= \tan^{-1}\left\{\left(\frac{1-0.1764}{0.42}\right)\left(\frac{10.8 * 16.8}{116.64 + 282.24}\right)\right\}$$

$$= \tan^{-1}\left\{\left(\frac{0.8236}{0.42}\right)\left(\frac{181.44}{398.88}\right)\right\}$$

$$\theta = \tan^{-1}\left(1.9610 * 0.4549\right)$$

$$= \tan^{-1}\left(0.8920\right)$$

$$\boxed{\theta = 41.7330^{\circ}}$$

Problem:- Calculate the correlation coefficient for the following data

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

$U = X - 68$

$V = Y - 72$

U (U-68)	V (U-72)	$U' = U - \bar{U}$	$V' = V - \bar{V}$	$(U')^2$	$(V')^2$	$U'V'$
-3	-5	-3	-2	9	4	6
-2	-4	-2	-1	4	1	2
-1	-7	-1	-4	1	16	4
-1	-4	-1	-1	1	1	1
0	0	0	3	0	9	0
1	0	1	3	1	9	3
2	-3	2	0	4	0	0
4	-1	4	2	16	4	8
$\Sigma U = 0$	$\Sigma V = -24$			$\Sigma U' = 36$	$\Sigma V' = 55$	

$\frac{36}{8} = 4.375$

$$r = \frac{\sum u'v'}{\sqrt{\sum (u')^2 \sum (v')^2}}$$

$$= \frac{24}{\sqrt{36 * 44}}$$

$$r = \frac{2\cancel{4}4}{6\sqrt{44}}$$

$$r = \frac{\cancel{4}2}{2\sqrt{11}}$$

$$r = 2/\sqrt{11}$$

$$\boxed{r = 0.6030}$$

Direct method:-

$U = X - 68$	$V = y - 69$	U^2	V^2	UV
-3	-2	9	4	6
-2	-1	4	1	2
-1	-4	1	16	4
-1	-1	1	1	1
0	3	0	9	0
1	3	1	9	3
2	0	4	0	0
4	2	16	4	8
		$\sum U^2 = 36$	$\sum V^2 = 44$	$\sum UV = 24$

$$r = \frac{n * \Sigma UV - \Sigma U \Sigma V}{\sqrt{n\Sigma U^2 - (\Sigma U)^2}\ \sqrt{n * \Sigma V^2 - (\Sigma V)^2}}$$

$$= \frac{8 * 24 - 0}{\sqrt{8(36) - 0}\ \sqrt{8(44) - 0}}$$

$$= \frac{24}{6 * 2\sqrt{11}}$$

$$= 2/\sqrt{11}$$

$$r = 0.6030$$

Rank Correlation :-

Problem :- Ten competitors in a music test were ranked by the three judges A, B & C in the following order

Ranks of A	1	6	5	10	3	2	4	9	7	8
Ranks of B	3	5	8	4	7	10	2	1	6	9
Ranks of C	6	4	9	8	1	2	3	10	5	7

Using the rank correlation method discuss which pair of judges has the nearest approch to common lickings in music

Ranks A X	Ranks B Y	Ranks C Z	$d_1 = x - y$	$d_2 = x - z$
1	3	6	-2	-5
6	5	4	1	2
5	8	9	-3	-4
10	4	8	6	2
3	7	1	-4	2
2	10	2	-6	0
4	2	3	2	1
9	1	10	8	-1
7	6	5	1	2
8	9	7	-1	1

$d_3=y-z$	d_1^2	d_2^2	d_3^2
-3	4	25	9
1	1	4	1
-1	9	16	1
-4	36	4	16
6	16	4	36
8	36	0	64
-1	4	1	1
9	64	1	81
1	1	4	1
2	1	1	4
	$\Sigma d_1 = 200$	$\Sigma d_2 = 60$	$\Sigma d_3 = 214$

$$P(x,y) = 1 - \left\{ \frac{6 * \Sigma d_1^2}{10(99)} \right\}$$

$$= 1 - \left\{ \frac{6 * 200}{10 * 99} \right\}$$

$$= 1 - \frac{40}{33}$$

$$P(x,y) = -0.2121$$

$$P(x,z) = 1 - \left\{ \frac{6 * 60}{10 \times 99_{33}} \right\}$$

$$= 1 - \left\{ \frac{2 \times 60}{3311} \right\}$$

$$\neq 1 - 0.06016$$

$$\neq$$

$$= 1 - \frac{4}{11}$$

$$= \frac{11-4}{11}$$

$$P(x,z) = 0.6364$$

$$\rho(Y,Z) = 1 - \frac{6 * 214}{10 * 99}$$

$$= 1 - \frac{214}{165}$$

$$= \frac{165 - 214}{165}$$

$$\rho(Y,Z) = -0.2970$$

∴ Inference $\rho(X,Z)$ is maximum we concluded that pairs of judges A and C as the nearest approach common likings in music.

Problem:- Calculate the correlation coefficient btw x & y for the following data

x	1	3	4	5	7	8	10
y	2	6	8	10	14	16	20

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	x^2	y^2	xy
1	2	-4.4286	-8.8571	19.6125	78.4482	39.224
3	6	-2.4286	-4.8571	5.8981	23.5914	11.796
4	8	-1.4286	-2.8571	2.0409	8.1630	4.081
5	10	-0.4286	-0.8571	0.1837	0.7346	0.3674
7	14	1.5714	3.1429	2.4693	9.8778	4.9388
8	16	2.5714	5.1429	6.6121	26.4494	13.2245
10	20	4.5714	9.1429	20.8977	83.5926	41.795
				$\Sigma x^2 =$	$\Sigma y^2 =$	$\Sigma xy =$
$\Sigma x = 38$	$\Sigma y = 76$			57.7143	230.857	115.428

$\Sigma \bar{x} = \frac{38}{7}$ $= 76/7$

$\Sigma \bar{x} = 5.4286$ 10.8571

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{115.4289}{\sqrt{57.7143 * 230.857}}$$

$$r = \frac{115.4289}{\sqrt{13323.7502}} = \frac{115.4289}{115.4286}$$

$$\boxed{r = 1.0000}$$

(2) The following table gives the no of blind per lakh of population in different age groups find out the correlation b/w age & blindness.

Age in years	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
no. of blind per lakh	55	67	100	111	150	200	300	500

x	x	y	$x = x - \bar{x}$	$y = y - \bar{y}$	x^2	y^2	xy
0-10	10	55	-30.625	-130.375	937.8906	16997.6406	3992.1344
10-20	15	67	-25.625	-118.375	656.6406	14012.6406	3033.3594
20-30	25	100	-15.625	-85.375	244.1406	7288.8906	1333.9844
30-40	35	111	-5.625	-74.375	31.6406	5531.6406	418.3594
40-50	45	150	4.375	-35.375	19.1406	1251.3906	-154.7656
50-60	55	200	14.375	14.625	206.6406	213.8906	210.2344
60-70	65	300	24.375	114.625	594.1406	13138.8906	2794.9844
70-80	75	500	34.375	314.625	1181.6406	98988.8906	10815.2344
	$\Sigma x = 325$	$\Sigma y = 1483$			$\Sigma x^2 = 3871.8748$	$\Sigma y^2 = 157423.8748$	$\Sigma xy = 19929.1252$
	$\bar{x} = 40.625$	$\bar{y} = 185.375$					

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{19929.1252}{\sqrt{3871.8748 * 157423.8748}} = \frac{19929.1252}{24688.5709} = 0.8072$$

$$\boxed{r = 0.8072}$$

2/11/2020
Monday

Partial Coefficient:-

Correlation

Sometimes the correlation b/w two variables x_1 & x_2 may be partly due to the

Coefficient of Partial Correlation:- Sometimes the correlation b/w two variables x_1 and x_2 may be partly due to the correlation of the third variable, x_3 with both x_1 and x_2 In such a situation, we may want to know, what the correlation b/w x_1 and x_2 would be if the effect of x_3 on each of x_1 and x_2 would be if the effect of x_3 on each of x_1 and x_2 were Eliminated this correlation is called partial correlation and the correlation coefficient b/w x_1 and x_2 after the linear effect of x_3 on each of them has been Eliminated is called the partial correlation coefficient

Thus the partial correlation coefficient b/w x_1 and x_2 usually denoted by

$$\rho_{12.3} = \frac{Cov(x_{1.3} | x_{2.3})}{\sqrt{Var(x_{1.3})\, Var(x_{2.3})}}$$

$$\boxed{\rho_{12.3} = \frac{Cov(x_{1.3}, x_{2.3})}{\sqrt{Var(x_{1.3})\, Var(x_{2.3})}}}$$

$$r_{12.3} = \frac{r_{12} - r_{13} * r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Note:- ① Similarly $r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}$

+

$$r_{23.1} = \frac{r_{23} - r_{21} * r_{31}}{\sqrt{(1-r_{21}^2)(1-r_{31}^2)}}$$

② If $r_{12.3}=0$, we have then $r_{12} = r_{13} * r_{23}$
If means that r_{12} will not be zero
If x_3 is correlated with both x_1 and x_2
Thus although x_1 and x_2 may be uncorrelated when Effect of x_3 is Eliminated, yet x_1 & x_2 may appear to be correlated because they carry the Effect of x_3 on them.

③ Partial correlation coefficient helps in deciding whether to include (or) not an additional independent variable in regression analysis

Problems:- Find the partial correlation $r_{12.3}$ from the following data

$r_{12}=0.77$; $r_{13}=0.72$, $r_{23}=0.52$

Sol^{n}:- given $r_{12}=0.77$

$r_{13}=0.72$

$r_{23}=0.52$

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{0.77 - (0.72)(0.52)}{\sqrt{(1-0.72^2)(1-0.52^2)}}$$

$$= \frac{0.77 - 0.3744}{\sqrt{(0.4816)(0.7296)}}$$

$$= \frac{0.77 - 0.3744}{0.3514}$$

$$= \frac{0.3956}{\sqrt{0.3514}} = \frac{0.3956}{0.5928}$$

$$\boxed{r_{12.3} = 0.6674}$$

In a trivariant distribution it is found $r_{12}=0.7$, $r_{13}=0.61$, $r_{23}=0.4$, Find the value of partial correlation cofficient of $r_{23.1}$, $r_{13.2}$, $r_{12.3}$.

$$r_{23.1} = \frac{r_{23} - r_{21} * r_{31}}{\sqrt{(1-r_{21}^2)(1-r_{31}^2)}}$$

$$= \frac{0.4 - (0.7)(0.61)}{\sqrt{(1-0.7^2)(1-0.61^2)}}$$

$$= -0.0477$$

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$= \frac{0.61 - (0.7)(0.4)}{\sqrt{(1-0.7^2)(1-0.4^2)}} \quad \frac{0.61-(0.7)(0.4)}{\sqrt{(1-0.7^2)(1-0.4^2}}$$

$r_{13.2} = -0.5041$

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{(0.7) - (0.61)(0.4)}{\sqrt{(1-0.61^2)(1-0.4^2)}} = 0.6279$$

Multiple correlation :- In studying the joint effect of a group of variables upon a variable not included in the group, such type of study is called multiple correlation.

For Example, the yield of the crop per acre say x_1 depends on the quality of seeds x_2, fertility of soil x_3, fertilizers used x_4, irrigation say x_5, weather conditions say x_6 and so on.

Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in the group. One study is that of multiple correlation.

Cofficient of multiple correlation :- In trivariate distribution in which each of the variables x_1, x_2 & x_3 has N observations, the multiple correlation cofficient of x_1 on x_2 & x_3 usually denoted by $R_{1.23}$ is the simple correlation b/w x_1 and the joint effect of x_2 and x_3 on x_1. In other words $R_{1.23}$ is the correlation cofficient b/w x_1 & its estimated value as given by the plane of regression of x_1 on x_2 & x_3

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{(1 - r_{23}^2)}$$

Similarly

$$R^2_{3.12} = \frac{r^2_{12} + r^2_{23} - 2 r_{12} r_{13} r_{23}}{(1 - r^2_{12})}$$

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{23}}$$

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{23} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{13}}$$

06/11/2020

Friday

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{23}}$$

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{12}}$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{13} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{13}}$$

Properties of multiple correlation:-

① Multiple correlation cofficient measures the closeness of the association b/w the observed values of a variable obtained from the multiple linear regression of that variable on other variables

② Mcc b/w observed values and the Expected values, when the Expected values are calculated from the linear relation of the variable determined by the method of least squares is always greater than that where expected values are calculated from any other linear combination of the variables

③ $0 \leq R_{1\cdot23} \leq 1$

④ If $R_{1\cdot23}=0$, then total & partial correlations involving X_1 are zero.

⑤ If $R_{1\cdot23}=1$ then association is perfect

⑥ $R_{1\cdot23} \geq r_{12}, r_{13}, r_{23}$

Problem:- Calculate $R_{1\cdot23}, R_{3\cdot12}, R_{2\cdot13}$ for the following data

$r_{12}=0.6$, $r_{13}=0.7$, $r_{23}=0.685$

Sol^n given that

$r_{12}=0.6$, $r_{13}=0.7$, $r_{23}=0.65$

$$R^2_{1\cdot23} = \frac{r_{12}^2 + r_{13}^2 - 2*r_{12}*r_{13}*r_{23}}{1 - r_{23}^2}$$

$\frac{608}{1155}$

$R^2_{1\cdot23} = (0.6)^2 + (0.7)^2 - 2*(0.6)(0.7)(0.65$

$R_{1.23} = 0.7255$

$$R^2_{3.12} = \frac{r_{13}^2 + r_{23}^2 - 2 r_{12} r_{13} * r_{23}}{1 - r_{12}^2}$$

$$= \frac{0.7^2 + 0.65^2 - 2(0.6 * 0.7 * 0.61)}{1 - 0.6^2}$$

$$= \frac{733}{1280}$$

$R^2_{3.12} = 0.5727$

$R_{3.12} = \sqrt{0.5727}$

$R_{3.12} = 0.7568$

$$R^2_{2.13} = \frac{r_{12}^2 + r_{23}^2 - 2 r_{12} r_{13} r_{23}}{1 - r_{13}^2}$$

$$= \frac{0.6^2 + 065^2 - 2 * 0.6 * 0.7 * 0.65}{1 - 0.7^2}$$

$= 0.4637$

$R^2_{2.13} = \sqrt{0.4637}$

$R^2_{2.13} = 0.6810$

problem:-

2 If $r_{12}=0.9$, $r_{13}=0.75$ and $r_{23}=0.7$ find $R_{1.23}$, $R_{3.12}$, $R_{2.13}$

$$R^2_{1.23} = \frac{r_{12}^2 + r_{13}^2 - 2*r_{12}*r_{13}*r_{23}}{1-r_{23}^2}$$

$$= \frac{0.9^2 + 0.75^2 - 2*0.9*0.75*0.7}{1-0.7^2}$$

$$= \frac{57}{68}$$

$$R^2_{1.23} = \frac{57}{68}$$

$$R_{1.23} = \sqrt{57/68}$$

$$R_{1.23} = 0.9156$$

$$R^2_{3.12} = \frac{r_{13}^2 + r_{12}^2 - 2*r_{12}*r_{13}*r_{23}}{1-r_{12}^2}$$

$$= \frac{0.75^2 + 0.9^2 - 2*0.9*0.75*0.7}{1-0.9^2}$$

$$R^2_{3.12} = 2.25$$

$$R_{3.12} = \sqrt{2.25}$$

$$R_{3.12} = 1.5$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{13} - 2 r_{12} r_{13} r_{23}}{1 - r^2_{13}}$$

$$= \frac{0.9^2 + 0.75^2 - (2 * 0.9 * 0.75 * 0.7)}{1 - 0.75^2}$$

$$= 0.9771$$

$$R_{2.13} = \sqrt{0.9771}$$

$$R_{2.13} = 0.9885$$

given that $r_{12} = 0.4$, $r_{13} = 0.5$, $r_{23} = 0.6$

find $r_{12.3}$ & $R_{12.3}$

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}}$$

$$= \frac{0.4 - (0.5)(0.6)}{\sqrt{(1 - 0.5^2)(1 - 0.6^2)}}$$

$$= \frac{\sqrt{3}}{12}$$

$$\boxed{r_{12.3} = 0.1443}$$

$$R^2_{1.23} = \frac{0.4^2 + 0.5^2 - 2 * 0.4 * 0.5 \times 0.6}{1 - 0.6^2}$$

$$R^2_{1.23} = 0.2656$$

$$R_{1.23} = 0.5154$$

Given that $r_{12}=0.6$, $r_{13}=0.72$ & $r_{23}=0.65$

find ~~$r_{1.23}$~~ $R_{1.23}$ & $r_{12.3}$

$$R^2_{1.23} = \frac{r^2_{12}+r^2_{13}-2*r_{12}*r_{13}*r_{23}}{1-r^2_{23}}$$

$$= \frac{0.6^2+0.72^2-2*0.6*0.72*0.65}{1-0.65^2}$$

$$R^2_{1.23} = 0.5486$$

$$R_{1.23} = 0.7407$$

$$r_{12.3} = \frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r^2_{13})(1-r^2_{23})}}$$

$$= \frac{0.6-0.72*0.65}{\sqrt{(1-0.7^2)(1-0.65^2}}$$

$$\boxed{r_{12.3} = 0.2432}$$

Correlation ratio:-

Correlation ratio 'η' is the appropriate measure of curvilinear relationship b/w the two variables

just as r measures the concentration of points about the curve of best fit. η measures the concentration of points about the curve of best fit.

Note:- If Regression is Linear, $\eta = r$ otherwise $|\eta| \geq |r|$

Measure of Correlation ratio:-

In Correlation we observe only one value of Y corresponding to ~~that~~ one value of X. There may be some cases the variable Y takes the value y_{ij} with frequencies f_{ij} where $j = 1, 2, \ldots n$ corresponding to the value of $x = x_i$ where $i = 1, 2, \ldots m$ These values can be shown in the following table.

y \ x	x_1	x_2	$\cdots$	x_i	$\cdots$	x_m	$n_{j.}$
y_1	$f_{11}(y_{11})$	$f_{21}(y_{21})$		$f_{i1}(y_{i1})$		$f_{m1}(y_{m1})$	n_1
y_2	$f_{12}(y_{12})$	$f_{22}(y_{22})$		$f_{i2}(y_{i2})$		$f_{m2}(y_{m2})$	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_j	$f_{1j}(y_{1j})$	$y_{2j}(y_{2j})$		$f_{ij}(y_{ij})$		$f_{mj}(y_{mj})$	$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	n_j
y_n	$f_{1n}(y_{1n})$	$f_{2n}(y_{2n})$		$f_{in}(y_{in})$		$f_{mn}(y_{mn})$	$\vdots$
							n_n
$n_{i.}$	n_1	$\cdots$	n_2	$\cdots$	n_i	$\cdots$ n_m	N

Here pair of observations (x_i, y_{ij}) with the frequencies f_{ij} so that

$$N = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$$

$$n_i = \sum_{j=1}^{n} f_{ij} \qquad j = 1, 2, \cdots n$$

$$n_j = \sum_{i=1}^{n} f_{ij}$$

$$T_i = \sum_{j=1}^{n} f_{ij} y_{ij}$$

$$T = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, y_{ij} = \sum_{i=1}^{m} T_i$$

Let us denote $\bar{y}_i$, $\bar{y}$ be the means of ith array and overall mean respectively

$$\bar{y}_i = \frac{\sum_{j=1}^{n} f_{ij}\, y_{ij}}{\sum_{j=1}^{n} f_{ij}} = \frac{T_i}{n_i}$$

$$\boxed{\bar{y}_i = \frac{T_i}{n_i}}$$

$$\bar{y} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} = \frac{T}{N}$$

$$\bar{y} = \frac{T}{N}$$

where $\boxed{T = \bar{y} \cdot N}$

The correlation ratio of y on x is denoted by η_{yx} defined as

$$\eta^2_{yx} = \frac{\sum_{i=1}^{m} \left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij})^2 - \left(\frac{T^2}{N}\right)}$$

Illarly the correlation ratio of x on y is denoted by η_{xy} & defined as

$$\eta^2_{xy} = \frac{\sum_{j=1}^{n} \frac{T_j^2}{n_j^2} - \left(\frac{T^2}{N}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij})^2 - \left(\frac{T^2}{N}\right)}$$

$$N = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} \; ; \; n_i = \sum_{j=1}^{n} f_{ij} \; ; \; n_j = \sum_{i=1}^{m} f_{ij}$$

$$T_i = \sum_{j=1}^{n} f_{ij} y_{ij} \; ; \; T = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} y_{ij} = \sum_{i=1}^{m} T_i$$

Properties of correlation ratio:-

1. Correlation ratio is independent of change of origin & scale

i.e,

$$\eta^2_{yx} = \eta^2_{xy} * \eta^2_{uv}$$

2. limits of correlation ratio are 0 to 1

i.e $0 \leq \eta^2_{yx} < 1$

Also

$$0 \leq \eta_{yx} < 1$$

9/11/2020

obtain correlation ratio of y on x to the following data

y \ x	10	50	20	25
7	3	2	–	–
9	–	1	4	6
11	–	3	4	2
13	2	1	5	–
15	–	6	–	1

The y on x correlation is

$$\eta^2_{yx} = \frac{\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \left(\frac{T^2}{N}\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, y_{ij}^2 - \frac{T^2}{N}}$$

where $N = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$

$$n_i = \sum_{j=1}^{n} f_{ij} \qquad T = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, y_{ij} = \sum_{i=1}^{m} T_i$$

$$n_j = \sum_{i=1}^{m} f_{ij}$$

$$T_i = \sum_{j=1}^{n} f_{ij}\, y_{ij}$$

$$\bar{y}_i = \frac{\sum_{j=1}^{n} f_{ij}\, y_{ij}}{\sum_{j=1}^{n} f_{ij}} = \frac{T_i}{n_i}$$

$$\boxed{T_i = n_i \cdot \bar{y}_i}$$

$$\bar{y} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} = \frac{T}{N}$$

$$\boxed{T = N \cdot \bar{y}}$$

13/11/2020 4488/4/1
Friday

y \ x	10	15	20	25	n_j	$f_{ij}y_{ij}$	$\sum_{i=1}^{m} f_{ij}y_{ij}$
7	3	2	—	—	5	$7(3)+7(2)=35$	$3\times7^2+2\times7^2=245$
9	—	1	4	6	11	$9(1)+9(4)+9(6)=99$	$1\times9^2+4\times9^2+6\times9^2=891$
11	—	3	4	2	9	$11(3)+11(4)+11(2)=99$	$3\times11^2+4\times11^2+2\times11^2=1089$
13	2	1	5	—	8	$13(2)+13(1)+13(5)=104$	$2\times13^2+1\times13^2+5\times13^2=1352$
15	—	6	—	1	7	$15(6)+15(1)=105$	$6\times15^2+1\times15^2=1572$
	5	13	13	9			
$f_{ij}y_{ij}$	$7(3)+13(2)$ 47	159	145	91	N=40	$T=\sum T_i=442$	$T=5152=\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}y_{ij}$
T_i^2	2209	25281	21025	8281	$\sum T_j^2=$ 56796		
$\frac{T_i^2}{n_i}$	$\frac{2209}{5}$ 441.8	1944.69	1617.3	920.11	4923.90		

$$\eta_{xy}^2=\frac{\sum T_i^2}{n_i}-\left(\frac{T^2}{N}\right)\Big/\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}y_{ij}^2-\left(\frac{T^2}{N}\right)$$

$$=\frac{4923.90-4884.1}{5152-4884.1}$$

$$\eta_{xy}^2=0.1486 \Rightarrow \eta_{xy}=\sqrt{0.1486}\Rightarrow 0.3855$$

*** Difference b/w Correlation & Regression:-

Correlation	Regression
1. Correlation means the relationship b/w two (or) more variables. It measures effect of one variable due to the change in the other variable.	1. Regression means stepping back to the mean value. It Expresses average relationship b/w two (or) more variables.
2. correlation need not imply cause and effect relationship b/w the variables under study	2. But Regression analysis indicates the cause & effect relationship. The variable constituting cause is taken as independent variable & the variable constituting the effect is taken as dependent variable.
3. Correlation analysis measure linear relationship only b/w the variables. Hence practical applications are very less	3. Regression analysis measures the linear & non linear relationships of the variables. Hence practical applications are more.
4. Sometimes correlation may be non-sense	4. There is no non-sense regression.
5. Correlation measures the direction & degree of linear relationship. It is symmetric. i.e, $r_{xy} = r_{yx}$. $r_{yx} = r_{xy}$ and it is immaterial whether the variables are dependent (or) independent	5. Regression analysis measures the functional relationship b/w the variables and it also used to estimate the value of dependent variable for any given independent variable. It identifies the nature of the variable i.e, which is independent and which is dependent. Regression coff are not symmetric i.e, $b_{yx} \neq b_{xy}$

H/W:-

① calculate rank correlation cofficient for the following data

X	10	15	12	17	13	16	24	14	22	20
Y	30	42	45	46	33	34	40	35	39	38

2. calculate the coffiaent of correlation for ranks from the following

(X, Y): (5,8), (10,3), (6,2), (3,9), (19,12), (5,3), (16,17), (12,18), (8,22), (2,12), (10,17), (19,20)

Analysis of Variance (ANOVA):-

Analysis of variance (ANOVA) is the seperation of variance ascriable to one group of causes from the variance ascribable to other group. ANOVA consists in the Estimation of amount of variation due to each of the independent factors (or) causes seperately and then comparing these Estimates due to assignable factors with the Estimates due to chance factors, the later being known as Experimental Error.

*** 1. ANOVA I one way classification
2. ANOVA II two way classification.

ie, $b_{yx} \neq b_{xy}$

ANOVA I way classification:-

Formulae:-

Rss = Raw sum of squares

$$Rss = \sum_{i=1}^{k} \sum_{j=1}^{h} x_{ij}^2$$

$$G = \text{Grand total} = \sum_{i=1}^{k} \sum_{j=1}^{h} x_{ij}$$

$$CF = \text{correction factor} = \frac{G^2}{N}$$

N = Total no of observations

Tss = Total sum of squares

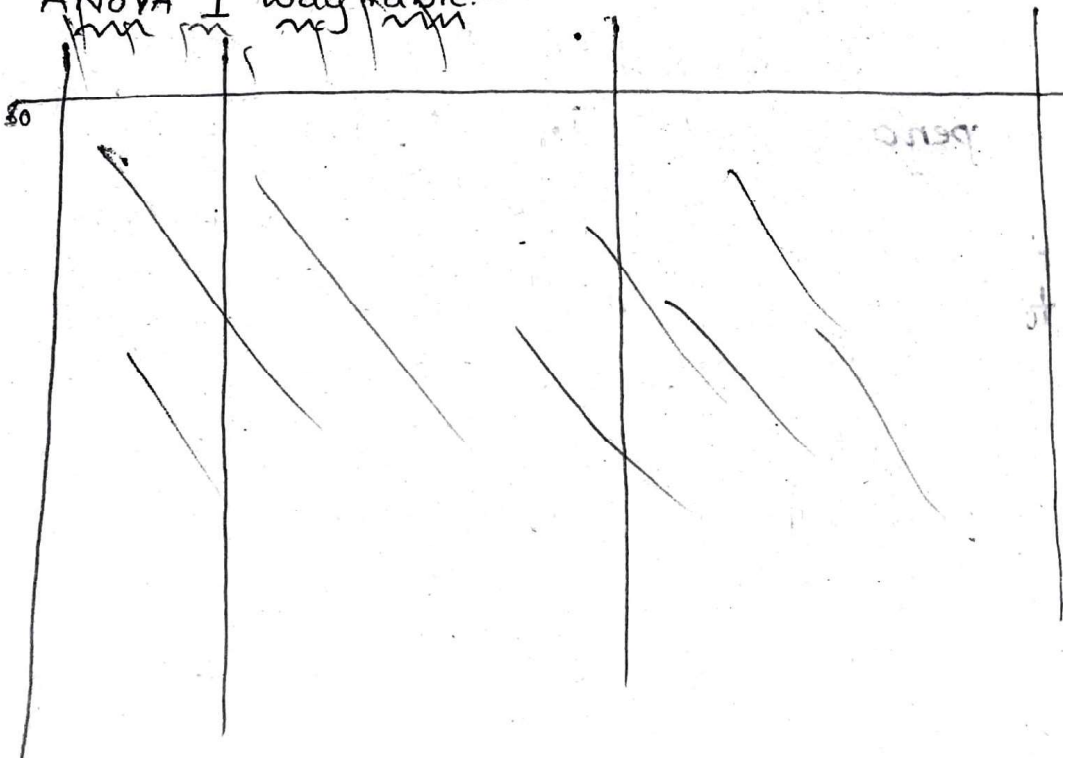
$$Tss = Rss - CF$$

$$SST = \text{Treatment sum of squares} = \sum_{i=1}^{k} \left(\frac{T_{i\bullet}^2}{n_i}\right) - CF$$

SSE = Error sum of squares (Sum of square due to error)

$$SSE = Tss - SST$$

ANOVA I way Table:-



ANOVA I Way Table:-

Source of variation	df degrees of freedom	Sum of Squares	mean sum of squares	Variance Ratio
Treatments	$k-1$	SST	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE} \sim F_{(k-1),(n-k)}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Tss	$n-1$			

Problem:- The following shows the times (in hours) of four batches of Electric labs perform an anlaysis of variance & hence show that a significance test doesn't reject their homogenity.

Batches	life of bulbs in hrs							
1	1600	1610	1650	1680	1700	1720	1800	—
2	1580	1640	1640	1700	1750	—	—	—
3	1460	1550	1600	1620	1640	1660	1740	1820
4	1510	1520	1530	1570	1600	1680	—	—

23/11/2020

1. From the following data, compute the coefficient of correlation b/w x & y

	x-series	y-series
no of items	15	15
Arthenetic mean	25	18
sum of squares of devation from mean	136	138

summation of product of deviations of x and y series from respective arthmetic mean = 122

	x	y
no of items	15	15
Arthematic mean	25	18
sum of squares of devation from mean	136	138

$\bar{x} = 25$
$\bar{y} = 18$

$x = 15, \quad y = 15$
$\bar{x} = 25 \quad \bar{y} = 18$
$\Sigma x^2 = 136 \quad \Sigma y^2 = 138$
$\Sigma xy = 122$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{122}{\sqrt{136 * 138}}$$

$$r = 0.8905$$

A sample of 12 fathers and there eldest son gave the following data about their height in inches

father 65 63 67 64 68 62 70 66 68 67 69 71

Sons 68 66 68 65 69 66 68 65 71 67 68 70

calculate the cofficient of rank correlation.

x	y	Rank of x_i	Rank of y_i	$x_i - y_i$	d_i^2
65	68	9	5.5	3.5	12.25
63	66	11	9	2	4
67	68	6.5	5.5	1	1
64	65	10	11.5	-1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	10	2	4
70	68	2	5.5	-3.5	12.25
66	65	8	11.5	-3.5	12.25
68	71	4.5	1	3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	-2.5	6.25
71	70	1	2	-1	1

$\frac{4+5}{2} = 9/2$

$\Sigma d_i^2 = 72$

$\frac{4+5+6+7}{4}$ $\frac{6+7}{2} = \frac{13}{2}$

CF in X series $= 67-2, 68-2$

$$\frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12}$$

$CF_X = 0.5 + 0.5 = 1$

$CF_y = 68 - 2 \quad 65 - 2$

$$CF_y = \frac{4(4^2-1)}{12} + \frac{2(2^2-1)}{12}$$

$$CF_y = \frac{4(15)}{12} + \frac{2(3)}{12}$$

$$= 5 + \frac{1}{2}$$

$$= \frac{11}{2}$$

$CF_y = 5.5 \quad 4.5 \quad 5.5$

$CF_x + CF_y = 1 + 5.5 = 6.5$

$$\rho = 1 - \left[\frac{6(72 + 6.5)}{12(143)}\right]$$

$$= 1 - \left[\frac{6 \times 78.5}{12 \times 143}\right]$$

$$= 1 - \left[\frac{471}{1716}\right]$$

$$= 1 - 0.2745$$

$$= 0.7255 \qquad (0.7220)$$

Ten competitors in a beauty contest are ranked by three judges as follows

judges	1	2	3	4	5	6	7	8	9	10
A	6	5	3	10	2	4	9	7	8	1
B	5	8	4	7	10	2	1	6	9	3
C	4	9	8	1	2	3	10	5	7	6

Discuss which pair has the nearest approach to common tastes of beauty

Rank A x	Rank B y	Rank C z	$d_1 = x - y$	$d_2 = x - z$	$d_3 = y - z$	d_1^2	d_2^2	d_3^2
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
3	4	8	-1	-5	-4	1	25	16
10	7	1	3	9	6	9	81	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
1	3	6	-2	-5	-3	4	25	9
						$\Sigma d_1^2 = 158$	$\Sigma d_2^2 = 179$	$\Sigma d_3^2 = 214$

$$P(X,Y) = 1 - \left\{ \frac{6 * \Sigma d_i^2}{n(n^2-1)} \right\}$$

$$= 1 - \left\{ \frac{6 * 158}{10(99)} \right\}$$

$$= 1 - \left\{ \frac{158}{165} \right\}$$

$$= 1 - \left\{ \frac{158}{165} \right\}$$

$$= 1 - 0.9576$$

$$= 0.0424$$

$$P(X,Z) = 1 - \left\{ \frac{6 * 179}{10(99)} \right\}$$

$$= 1 - \left\{ \frac{1074}{990} \right\}$$

$$P(X,Z) = 0.0848$$

$$P(Y,Z) = 1 - \left\{ \frac{6 * 214}{10 * 99} \right\}$$

$$= 1 - \left\{ \frac{1284}{990} \right\}$$

$$= 1 - 1.2970$$

$$P(Y,Z) = 0.2970$$

27/11/2020

Marks of 6 students of a class in paper I and paper II

1) both Regression coefficients
2) both Regression lines
3) Re. correlation coefficient

paper I 45 55 66 75 85 100

paper II 56 55 45 65 62 71

x	y	$x-\bar{x}$	$y-\bar{y}$	x^2	y^2	xy
45	56	-11-26	-3	676	9	78
55	55	-6-16	-4	256	16	64
6.6	45	27 -5	-14	25	196	70
75	65	16 4	6	16	36	24
85	62	23 14	3	196	9	42
100	71	29 29	12	841	144	348
Σx = 426	Σy = 354			Σx^2 = 2010	Σy^2 = 410	Σxy = 626
71	59					

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

$$= \frac{626}{\sqrt{2010 * 410}}$$

$$r = 0.6896$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma x^2}$$

$$= \sqrt{\frac{1}{6} * 2010}$$

$$\sigma_x = 18.3030$$

$$\sigma_y = \sqrt{\frac{1}{6} * 410}$$

$$\sigma_y = 8.2664$$

(71)

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$= 0.6896 * \frac{8.2664}{18.3030}$$

$$= 0.3114$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{628}{2010}$$

$$= \frac{314}{1005}$$

$$= 0.3124$$

~~b_{yx}~~

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$= 0.6896 * \frac{18.3030}{8.2664}$$

$$b_{xy} = 1.5269$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2}$$

$$= \frac{628}{410}$$

$$b_{xy} = 1.5317$$

Regression line y on x:-

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$y - 59 = 0.3124(x - 71)$$

$$y - 59 = 0.3124x - 22.1804$$

$$y - 59 = 0.3124x - 22.1804$$

$$y = 0.3124x + 36.8196$$

Regression line on x on y:-

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$x - 71 = 1.5317(y - 59)$$

$$x - 71 = 1.5317y - 90.3703$$

$$x = 1.5317y \mp 19.3703$$

$$x = 1.5317 * 0.49 \mp 19.3703$$

$$= 0.7505 \mp 19.3703$$

~~$x = 20.1208$~~

$$\boxed{x = -18.6198}$$

y ~~for~~ $x = 0.2$

$$y = 0.3124 * 0.2 + 36.8196$$

$$= 0.06248 + 36.8196$$

$$\boxed{y = 36.8820}$$

2. regression line of y on x on y respectively are

$$2x - 3y = -8$$

$$5x - y = 6$$

then find

(i) mean values of x and y

(ii) cofficient of correlation b/w x and y

(iii) standard deviation of Y for given variance of $x = 5$

eqn ① as y on x & eqn ② is x on y (7)

$2x - 3y = -8$

$-3y = -8 + 2x$

$-3y = 2x - 8$

$y = +\frac{2x}{3} + \frac{8}{3}$

$b_{yx} = +\frac{2}{3}$

$2x - 3y = -8$

$2x = 3y - 8$

$x = \frac{3}{2}y - \frac{8}{2}$

$b_{xy} = 3/2$

eqn ② is x on y

$5x - y = 6$

$5x = y + 6$

$x = \frac{y}{5} + \frac{6}{5}$

$b_{xy} = 1/5$

$-y = -5x + 6$

$y = 5x - 6$

$b_{yx} = 5$

$r = \sqrt{b_{xy} * b_{yx}}$

$= \sqrt{(+2/3)(1/5)}$

$= 0.3651$

$r = \sqrt{3/2 * 5}$

$= \sqrt{15/2}$

$r = 2.7386$

$2x - 3y = -8$

$5x - y = 6$

$x = 2$

$y = 4$

$r = \sqrt{b_{xy} * b_{yx}}$

$= \sqrt{\left(\frac{2}{3}\right) * 1/5}$

$= \sqrt{\frac{2}{15}}$

$r = 0.3651$

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$\frac{2}{3} = 0.3651\left(\frac{\sigma_y}{5}\right)$

$\frac{2}{3} \times 5 = 0.3651\,(\sigma_y)$

$\frac{10}{3 \times 0.3651} = \sigma_y$

$\frac{2}{3} = \frac{0.3651}{2.2360}\,(\sigma_y)$

$\frac{4.4721}{1.0953} = \sigma_y$

$\sigma_x^2 = 5$

$\sigma_x =$

~~$\sigma_y = b_{yx} \times b_{yx}$~~

$\sigma_y = \frac{\sigma_x * b_{yx}}{r}$

$\sigma \cdot b_{yx}$

1. Calculate the karl pearson's cofficient of correlation b/w price and demand for the following data

price	17	18	19	20	22	24	26	28	30
demand	40	38	35	30	28	25	22	21	20

x	y	$x-\bar{x}$	$y-\bar{y}$	x^2	y^2	xy
17	40	−5.6667	11.2222	32.1115	125.9378	
18	38	−4.6667	9.2222	21.778	85.0490	
19	35	−3.6667	6.2222	13.4447	38.7158	
20	30	−2.6667	1.2222	7.1113	1.4938	
22	28	−0.6667	−0.7778	0.4445	0.6050	
24	25	1.3333	−3.7778	1.7777	14.2718	
26	22	3.3333	−6.7778	11.1109	45.9384	
28	21	5.3333	−7.7778	28.4441	60.4926	
30	20	7.3333	−8.7778	53.7773	53.77	

$\Sigma x = 204$ $\Sigma y = 259$ $\Sigma x^2 = 170$ $\Sigma y^2 = 381.0$

$= 28.7778$

$= 22.6667$

-63.5928 $\Sigma xy = -265.6663$

-43.0372

-22.8149 $r = \frac{-265.6663}{\sqrt{170 * 381.0739}}$

-3.2592

0.5186

-5.0369

-22.5924

-41.4813

-4.3702

30/11/2020:
Monday

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (null hypothesis)

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ (Alternative hypothesis)

Coding method:- Subtract Each number from 1600 & divide by 10.

									T_j
1	0	-1	-5	-8	-10	-12	-20	—	-56
2	+102	-4	-4	-10	-15	—	—	—	-31
3	+154	5	0	-2	-4	-6	-14	-22	-29
4	9	8	7	3	0	-8	—	—	19

$G = -97$

$N = 26$

$RSS = 2319$

$G = -97$ (grand total)

$G = \sum\sum x_{ij} = -97$

Correction factor $= \frac{G^2}{N} = \frac{9409}{26}$

Correction factor $= 361.8846$.

Total sum of squares $= TSS$

$= RSS - CF$

$= 2319 - 361.8846$

$= 1957.1154$

~~Sum of squares due to Total~~

Sum of squares due to Treatments

$$= SST = \sum_{i=1}^{4} \left(\frac{T_{i.}^2}{n_i} \right) - CF$$

$$= \frac{(-56)^2}{7} + \frac{(-31)^2}{5} + \frac{(-29)^2}{8} + \frac{(19)^2}{6} - CF$$

$$= \frac{96659}{120} - 361.8846$$

$$= 805.4917 - 361.8846$$

$$SST = 443.6071$$

Sum of squares due to Errors

$$SSE = TSS - SST$$

$$= 1957.1154$$

$$SSE = 1513.5083$$

ANOVA I way table:-

Source of variation	degree of freedom	Sum of squares	mean sum of squares	Variance ratio
Treatments (b/w batches)	$4-1=3$ $K-1$	$SST = 443.6071$	$MST = \frac{SST}{K-1}$ $\frac{443.6071}{3}$ 147.8690	$F = \frac{MST}{MSE}$ $= \frac{147.8690}{68.7952}$ $= 2.1494$
Error	$N-K$ $26-4=22$	SSE 1513.5083	$MSE = \frac{SSE}{22}$ 68.7952	
Total	$N-1$ $26-1=25$			

* The variance ratio the numerator should be always greater than denominator
If the condition then reverse the fraction

Table value of F at 5% level of significance at (3,22) degrees of freedom is 3.0491

Calculated value is < Table value than. accept H_0.

problem:- 4 sales men were posted in different areas by a company. The no. of units of commodity X sold by them are as follows. Is there significant difference the performance of salesmen?

A	20	23	28	29
B	25	32	30	21
C	23	28	35	18
D	15	21	19	25

$H_0 \Rightarrow \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1 \Rightarrow \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

Coding method:- Subtract Each number from 28

					T_i	x_{ij}^2
A	8	5	0	-1	12	90
B	3	-4	-2	7	4	78
C	5	0	-7	10	8	174
D	13	7	9	3	32	308
					G=56	RSS=650

N=16, RSS=650, $G = \sum\sum x_{ij} = 650$

Correction factor $= \frac{G^2}{N} = \frac{56^2}{16}$

$= \frac{3136}{16}$ 196

Correction factor $= 6601.5625$

Total sum of squares $=$ Tss $=$ Rss $-$ CF

$= 650 - 6601.5625$

$= 650 - 196$

$= 454$

Sum of squares due to treatments

$= SST = \sum_{i=1}^{4} \left(\frac{T_i^2}{n_i}\right) - CF$

$\frac{(12)^2}{4} + \frac{(4)^2}{4} + \frac{(8)^2}{4} + \frac{(32)^2}{4} - 454$

$SST = 312 - 454$

$SST = -142$ (116)

Sum of squares due to Errors:-

$SSE = Tss - SST$

$= 454 - (-142)$ (454 $-$ 116)

$= 454 + 142$

$SSE = 596$ (338

Table value of F at 5% level of significance at 3, 12 df is 3.4903

ANOVA I way table :-

source of Variation	degree of freedom	sum of squares	mean of sum squares	Variance ratio
Treatments (batches)	$k-1 \Rightarrow 4-1 = 3$	SST = 142	$MST = \frac{SST}{k-1}$ $\frac{142}{3} = 47.3333$	$F = \frac{47.3333}{49.6667} = 0.9531$
Error	$N-k$ $16-4 = 12$	SSE = 596	$MSE = \frac{SSE}{N-k}$ $\frac{596}{12}$ 49.6667	
Total	$N-1$ $16-1 = 15$	TSS		

$0.9531 < 3.4903$ then accept H_0.

4/12/2020

ANOVA II way :-

Rows → k

columns → h

Formulae:- $TSS = RSS - CF$

$$CF = \frac{G^2}{N}.$$

$$SST = \frac{1}{h}\sum T_j^2 - CF \quad (h\text{-columns})$$

$$SSV = \frac{1}{k}\sum T_j^2 - CF$$

$$G = \sum_{i=1}^{k}\sum y_{ij} \qquad T_i = \sum_{i=1}^{k} y_{ij}$$

$$T_j = \sum_{j=1}^{h} y_{ij}$$

$$SSE = TSS - SST - SSV$$

ANOVA II way Table:-

Source of Variance	d.f	Sum of squares	Mean Sum of squares	Variance Ratio
Treatments	$k-1$	SST	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE} \sim F_{(k-1),(k-1)(h-1)}$
varieties	$h-1$	SSV	$MSV = \frac{SSV}{h-1}$	$F = \frac{MSV}{MSE} \sim F_{(h-1),(k-1)(h-1)}$
Errors	$(k-1)*(h-1)$	SSE	$MSE = \frac{SSE}{(k-1)(h-1)}$	
Total	$hk-1$			

numerator is greater than denominator

problem:- A tea company appoints 4 salesmen A, B, C and D & observes that sales in three seasons — summer, winter & monsoon of the figures (in lakhs) are given in the following table

seasons	A	B	C	D	seasons total
summer	36	36	21	35	128
winter	28	29	31	32	120
monsoon	26	28	29	29	112
salesmen Total	90	93	81	96	360

$H_{01} \Rightarrow \mu_1 = \mu_2 = \mu_3$

$H_{11} \Rightarrow \mu_1 \neq \mu_2 \neq \mu_3$

$H_{02}: \mu_1' = \mu_2' = \mu_3' = \mu_4'$

$H_{12}: \mu_1' \neq \mu_2' \neq \mu_3' \neq \mu_4'$

Coding method:- Subtract each value from 29

					$T_{i.}$	$T_{i.}^2$
	-7	-7	8	-6	-12	144
	1	0	-2	-3	-4	16
	3	1	0	0	4	16
$T_{.j}$	-3	-6	6	-9	$G = -12$	$\Sigma T_i = 176$
$T_{.j}^2$	9	36	36	81	~~$\Sigma\Sigma y_{ij}$~~ $= 162$	

$\Sigma T_j^2 = 162$

$N = 12$

$G = -12$

$$CF = \frac{G^2}{N} = \frac{(-12)^2}{12} = 12$$

$$RSS = (-7)^2 + (-7)^2 + 8^2 + (-6)^2 + (1)^2 + 0^2 + (-2)^2 + (-3)^2 + (3)^2 + 1^2 + 0^2 + 0^2$$

$$= 222$$

$$SST = \frac{1}{h} \sum T_{i.}^2 - CF$$

$$= \frac{1}{4}(176) - 12$$

$$SST = 32$$

$$SSV = \frac{1}{k} \sum T_{.j}^2 - CF$$

$$= \frac{1}{3}(162) - 12$$

$$= 42$$

$$TSS = RSS - CF$$

$$= 222 - 12$$

$$TSS = 210$$

$$SSE = TSS - SST - SSV$$

$$= 210 - 32 - 42$$

$$SSE = 136$$

ANOVA II-way table

Source of variance	d.f	Sum of squares	mean sum of squares	F.ratio
Treatments	3-1=2	32=SST	MST=32/2=16	F=MST/MSE = 22.6667/16 =1.4167
varieties	4-1-3	SSV=42	MSV=42/3= =14	F=MSV/MSE = 22.6667/14 =1.6191
Error	(3-1)(4-1) 2*3=6	SSE=136	MSE=136/6 =22.6667	
Total	hk-1 =12-1 =11			

Since the F ratios are < 1 accept Ho.

1.4167 < 19.33 calculated value < Table value
Accept Ho1

1.619 < 8.94 calculated value < Table value
Accept Ho2

	df		
T	2	F at (2,6)	Numerator > denominator.
V	3	F at (3,6)	
E	6	F at (6,2)	Numerator < denominator.
		F at (6,3)	

There is no significant difference b/w the seasons

There is no significant difference b/w the performance of salemen.

Problem:- The following table represents the sales (Rs 1000) per month of 3 brands of soaps allocated among three cities

Brands	cities A	B	C
I	12	48	30
II	42	54	57
III	9	42	21

Test whether ① mean sales of three brands are equal ② mean sales of soaps in each city are equal

Brands	A	B	C	$T_{i.}$	$T_{i.}^2$
I	30	-6	12	36	1296
II	0	-12	-15	-27	729
III	33	0	21	54	2916
~~sales total~~ T_j	63	-18	18	$G=63$	$\Sigma T_{i.}^2 = 4941$
T_j^2	3969	324	324	$\Sigma T_j^2 = 4617$	

coding method:- Subtract with 42 (each number)

$H_{01} \Rightarrow \mu_1 = \mu_2 = \mu_3$

$H_{11} \Rightarrow \mu_1 \neq \mu_2 \neq \mu_3$

$H_{02} \Rightarrow \mu_1' = \mu_2' = \mu_3'$

$H_{12} \Rightarrow \mu_1' \neq \mu_2' \neq \mu_3'$

$N = 9$

$G = 63$

$CF = \frac{G^2}{N} = \frac{(63)^2}{9} = 441$

$$RSS = (30)^2 + (-6)^2 + (12)^2 + (0)^2 + (-12)^2 + (-15)^2 + (33)^2 + (0)^2 + (21)^2$$

$$RSS = 2979$$

Rows $\Rightarrow 3$
Columns $\Rightarrow 3$

$$SST = \frac{1}{h} \sum T_{i.}^2 - CF$$

$$= \frac{1}{3}(4941) - 441$$

$$= 1647 - 441$$

$$SST = 1206$$

$$SSV = \frac{1}{3} \sum T_{.j}^2 - CF$$

$$= \frac{1}{3}(4617) - 441$$

$$= 1539 - 441$$

$$SSV = 1098$$

$$TSS = RSS - CF$$

$$TSS = 2538$$

$$SSE = TSS - SST - SSV$$

$$= 2538 - 1206 - 1098$$

$$SSE = 234$$

ANOVA II way Table

Source of Variance	d.f	Sum of Squares	mean Sum of squares	F. ratio
Treatments	3-1=2	SST=1206	MST=$\frac{1206}{2}$=603	10.30 MST/MSE = 5.1538
Varities	3-1=2	SSV=1098	MSV=$\frac{1098}{2}$=549	MSV/MSE = 4.692
Error	2*2 =4	SSE=234	MSE=$\frac{234}{4}$=58.5	
Total	hk-1 9-1=8			

(2,4) ⇒ 19.25

(2,4) ⇒ 19.25

Tab val of Fat (2,4) df is at 5% los is 6.94

cal val > Tab val

Reject H_{01}

Tab val of Fat (2,4) df at 5% los is 6.94.

cal val > Tab val

Reject H_{02}

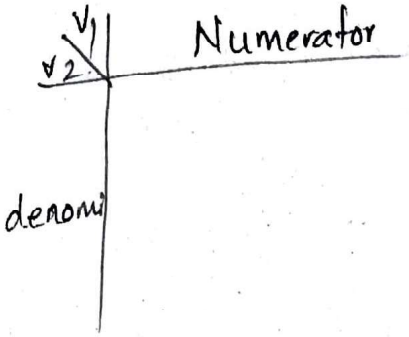
Inference:- ① The mean sales of three brands significantly differ

② The mean sales of soap in each city are not equal

$V_1 \leftarrow D$

$V_2 \rightarrow V_2$

(2, 4)



7/12/2020

① A trucking company wishes to test the average life of each of the four brands of tyres the company uses all brands. On randomly selected trucks. The records showing the lifes (1000 of miles) of tyres are given in the table.

Test the hypothesis that the average life of each brand of tyres is the same.

Assume $\alpha = 0.01$ (level of significance (If, we have) to see)

Brand I	Brand II	Brand III	Brand IV
20	19	21	15
23	15	19	17
18	17	20	16
17	20	17	18
—	16	16	—

$H_0 \Rightarrow \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_1 \Rightarrow \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

						T_i
Brand I	20	23	18	17	—	78
Brand II	19	15	17	20	16	87
Brand III	21	19	20	17	16	93
Brand IV	15	17	16	18	—	

coding method :- subtract each no. from 17

					T_i	x_{ij}^2
-3	-6	-1	0	—	-10	46
-2	2	0	-3	1	-2	18
-4	-2	-3	0	1	-8	30
2	0	1	-1	0	2	6
					$G = 18$	$RSS = 100$

$N = 18$

$G = -18$

$RSS = 100$

$CF = \frac{G^2}{N} = \frac{(-18)^2}{18} = 18$

$TSS = RSS - CF$

$= 100 - 18$

$TSS = 82$

$$SST = \left(\frac{100}{4}\right) + \frac{4}{5} + \frac{64}{5} + \frac{4}{4} \quad (90)$$

$$= \frac{198}{5} - 18$$

$$= \frac{108}{5}$$

$$SST = 21.6$$

$$SSE = TSS - SST$$

$$= 82 - 21.6$$

$$SSE = 60.4$$

Source of variation	d.f	Sum of squares	mean sum of squares	Variance ratio
Treatments (Brands)	$r-1$ $4-1=3$	$SST = 21.6$	$MST = \frac{21.6}{3} = 7.2$	$F = \frac{MST}{MSE} = \frac{7.2}{4.3143} = 1.6689$
Error	$N-K$ $18-4=14$	$SSE = 60.4$	$MSE = \frac{60.4}{14} = 4.3143$	
	$18-1=17$	$TSS = 82$		

$(3,14) \Rightarrow$ ~~16.69~~ 5.56

Table value of F at 1% level of significance at (3,14) df is

Calculated value < Table value Accept H_0.

Inference:- average life of Each brand of tyres is the same.

H/w

(2) Setup an ANOVA table for the following per hectare yield in (100 kg) for the three varities of wheat grown

Varities of wheat	I	II	III	IV
A1	6	7	3	8
A2	5	5	3	7
A3	5	4	3	4.

$H_{01} \Rightarrow \mu_1 = \mu_2 = \mu_3$.

$H_{11} \Rightarrow \mu_1 \neq \mu_2 \neq \mu_3$

$H_{02} \Rightarrow \mu_1' = \mu_2' = \mu_3' = \mu_4'$

$H_{12} \Rightarrow \mu_1' \neq \mu_2' \neq \mu_3' \neq \mu_4'$.

#Two

(2) Setup an ANOVA table for the following per hectare yield in (100 kg) for the three varities of wheat grown

varities of wheat	I	II	III	IV
A1	6	7	3	8
A2	5	5	3	7
A3	5	4	3	4

$H_{01} \Rightarrow \mu_1 = \mu_2 = \mu_3$

$H_{11} \Rightarrow \mu_1 \neq \mu_2 \neq \mu_3$

$H_{02} \Rightarrow \mu_1' = \mu_2' = \mu_3' = \mu_4'$

$H_{12} \Rightarrow \mu_1' \neq \mu_2' \neq \mu_3' \neq \mu_4'$

varities of wheat	I	II	III	IV	T_i	T_{ij}^2
A1	-1	-2	2	-3	-4	16
A2	0	0	2	-2	0	8
A3	0	1	2	1	4	16
T_j	-1	-1	6	-4	G=0	32 = RSS
T_j^2	1	1	36	16		

$N = 12$

$RSS = 32$

$CF = 0$

$TSS = 32$

$SSV = \frac{1}{3}(54) - 0 = 18$

$SST = \frac{16}{4} + \frac{16}{4}$

$SST = 8$

$SSE = TSS - SST - SSV$

$= 32 - 18 - 8$

$= 24 - 6$

source of Variation	d.f	Sum of squares	Mean sum of squares	Variance Ratio
Treatments	$3-1=2$	$SST=8$	$MST=\frac{8}{2}=4$	$\frac{MST}{MSE}=\frac{4}{1}=4$
varieties	$4-1=3$	$SSV=18$	$MSV=\frac{18}{3}=6$	$\frac{MSV}{MSE}=\frac{6}{1}=6$
Error	$2\times3=6$	$SSE=6$	$MSE=\frac{6}{6}=1$	
total	$12-1=11$			

$(2,6)$ $(3,6)$

Table value of F at 5% los at (2,6)df is

5.14 calval < Tabval

Accept H_{01}

Table value of F at 5% los at (3,6)df is

4.71 calval > Tabval Rejeet H_{02}

Inference:- (1) There is no significant difference b/w the three varities of wheat

(2) There is significant difference b/w the four yields of land.

11/12/2020
Friday

Estimation :- The theory statistical inference can be divided into two major areas
① Estimation of parameters
② Testing of hypotheis

Parameter :- A statistical constant derived from the population values
Ex:- Population mean, μ
Population standard deviation, σ

Statistic :- A statistical constant derived from the sample values is called a 'statistics'

Estimation :- ~~Procedure~~ It is a procedure of Estimating a population parameter by using sample information (or) Observation

1) Point Estimation
2) Interval Estimation

Point Estimate :- A point Estimate of some population parameter θ is a single numerical value

Point Estimator :- A point Estimator is a statistic for Estimating a population parameter θ & is denoted by $\hat{\theta}$

$\mu = (25) \rightarrow$ Estimate

$\hat{\theta} \leftarrow \bar{x} = 25.2 \rightarrow$ point Estimator
($\bar{x}$ ← sample mean)

Properties of Estimators:-

A estimator is said to be a good estimator if it is

1. unbiased
2. consistant
3. Efficient
4. sufficient.

(or)

1. unbiasedness
2. Consistency
3. Efficiency
4. sufficiency

Unbiased Estimator:- A statistic $\hat{\theta}$ is said to be an unbiased estimator (or) its value an ubiased estimate if & only if the mean of the sampling distribution of the Estimator equals θ i.e;

$$\boxed{\text{Mean of } \hat{\theta} = E(\hat{\theta}) = \theta}$$

$$\boxed{E(\hat{\theta}) = \theta}$$

r.v
discrete cont
$E(x) = \sum x f(x)$ / $E(x) = \int_{-\infty}^{\infty} x f(x)\,dx$

Sampling distribution:- Sample size = 2 = n

Sample no = N

N = 5
n = 2

1
2
3
4
5

If the Estimator is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the bias of the Estimate $\hat{\theta}$

when the ~~esti~~ Estimator is unbiased,

$$\boxed{E(\hat{\theta}) - \theta = 0}$$

i.e., the bias is zero.

positive bias :- If $E(\hat{\theta}) > \theta$, then $\hat{\theta}$ is called positively biased

negative bias :- If $E(\hat{\theta}) < \theta$, then $\hat{\theta}$ is called negatively biased.

(1) problem :- If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased Estimators of θ, then $w_1\hat{\theta}_1 + w_2\hat{\theta}_2$ is an unbiased Estimator of θ when $w_1 + w_2 = 1$

Sol^n :- given that

$\hat{\theta}_1$ & $\hat{\theta}_2$ are unbiased Estimators of θ

i.e., $E(\hat{\theta}_1) = \theta \longrightarrow (1)$

$E(\hat{\theta}_2) = \theta \longrightarrow (2)$

Now, $E(w_1\hat{\theta}_1 + w_2\hat{\theta}_2) = w_1 E(\hat{\theta}_1) + w_2 E(\hat{\theta}_2)$

$= w_1\theta + w_2\theta$ (from eq^n (1) & (2))

$= (w_1 + w_2)(\theta)$ $(w_1 + w_2 = 1)$

$E(w_1\hat{\theta}_1 + w_2\hat{\theta}_2) = \theta$

$\therefore$ $w_1\hat{\theta}_1 + w_2\hat{\theta}_2$ is an unbiased Estimator of θ.

Theorem:- Let $x_1, x_2, \ldots x_n$ be a random sample drawn from a given population with the mean μ & Variance σ^2 show that the sample mean is an unbiased Estimator of population mean μ

i.e, $E(\bar{x}) = \mu$.

proof:- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Taking Expectations on both sides

$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n} E(x_1 + x_2 + \ldots + x_n)$$

$$= \frac{1}{n} \{E(x_1) + E(x_2) + \ldots + E(x_n)\}$$

$$= \frac{1}{n} \{\mu + \mu + \ldots + \mu\} \quad [\because E(x_i) = \mu]$$

$$= \frac{1}{n} [n\mu]$$

$$= \frac{1}{n} [n\mu]$$

$$\boxed{E(\bar{x}) = \mu}$$

Theorem:- for a random sample of size n, $x_1, x_2, \ldots x_n$ taken from a finite population $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is not an unbiased Estimator of σ^2 but $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is an unbiased Estimator of σ^2

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

MVUE (minimum variance unbiased estimate)
(or) most efficient unbiased Estimator

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are the two unbiased estimators of θ if and if σ_1^2 and σ_2^2 are variances of their sampling distributions and $\sigma_1^2 < \sigma_2^2$, $\hat{\theta}_1$ is said to be more unbiased estimator of θ

Note:- If $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_k$ are unbiased Estimators of the parameter θ

$\sigma_1^2, \sigma_2^2, \ldots \sigma_k^2$ are variances of there sampling distributions then $\hat{\theta}_1$ is the most efficient estimator of θ if

$\theta \; \sigma_1^2 < \sigma_2^2 < \ldots < \sigma_k^2$

e = efficaency of an estimator

$e = \frac{\text{Variance } \theta_1}{\text{Variance } \theta_i'}$ where $i = 2, 3, \ldots k$ $(e < 1)$

$e = \frac{\text{var } \theta}{\text{var } \theta_1'}$

$e = \frac{\text{var } \theta_1}{\text{var } \theta_i}$

Consistency:- An estimator $T_n = T(x_1, x_2, \ldots x_n)$ based on a random sample of size n, is (gamma) Said to be a consistent estimator of $\gamma(\theta)$, $\theta \in \Theta$, $\Theta \rightarrow$ parameter space, if T_n converges to $\gamma(\theta)$

i.e, if $T_n \xrightarrow{P} \gamma(\theta)$ as $n \to \infty$

In other words T_n is a consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0$, $\eta > 0$, there exists a positive integer $n \geq m(\varepsilon, \eta)$ such that $P\{|T_n - \gamma(\theta)| < \varepsilon\} \to 1$ as $n \to \infty \Rightarrow P\{|T_n - \gamma(\theta)| < \varepsilon\} \geq 1 - \eta \ \forall n \geq n$ where n is some very large value n.

$\gamma, \theta \to$ parameter

$T_n \to$ estimator

18/12/2020

Sufficiency :- An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter

If $T = t(x_1, x_2, \ldots x_n)$ is an estimator of a parameter θ based on a sample $x_1, x_2, \ldots x_n$ of size n from the population with density $f(x, \theta)$ such that the conditional distribution of $x_1, x_2, \ldots x_n$ given T, is independent of θ; then t is sufficient estimator for θ

Methods of Estimation:-

1. maximum likehood method
2. method of moments
3. method of minimum chi-square
4. method of minimum variance
5. method of least squares
6. method of inverse probability

Maximum likelihood Estimation:- It was formulated by C.F Gauss as a general method of estimation was introduced by professor by R.A fisher

likelihood fn:- Let $x_1, x_2 \cdots x_n$ be a random sample of size n from a population with density fn $f(x,\theta)$ Then the likelihood fn of the sample values $x_1, x_2, \cdots x_n$ usually denoted by $L = L(\theta)$ is their joint density fn

$$L = f(x_1,\theta) f(x_2,\theta) \cdots f(x_n,\theta)$$

$$= \prod_{i=1}^{n} (f(x_i,\theta))$$

L gives the relative likelihood that the random variables assume a particular set of values $x_1, x_2 \cdots x_n$. For a given sample $x_1, x_2 \cdots x_n$ L becomes a fn of the variable θ, the parameter

The principle of maximum likelihood estimation consists in finding a estimator for the unknown parameter $\theta = (\theta_1, \theta_2, \cdots \theta_k)$ which maximises the likelihood function $L(\theta)$ for variation in parameter

If there exists a fn $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots x_n)$ of the sample values which maximises L for variations in θ, then $\hat{\theta}$ is to be taken as an estimator of θ

$\hat{\theta}$ is usually called maximum likelihood estimator

Thus $\hat{\theta}$ is the sln of

$$\frac{\partial L}{\partial \theta} = 0 \quad \& \quad \frac{\partial^2 L}{\partial \theta^2} < 0$$

$\hookrightarrow (1)$ $\hookrightarrow (2)$

since $L > 0$ & $\log L$ is nondecreasing fn of L, L & $\log L$ attain their extreme values (maximal minima) at the same value of $\hat{\theta}$

(1) can be rewritten as

$$\frac{1}{L} \frac{\partial L}{\partial \theta} = 0$$

This is a convinient form from practical view

If θ is a vector valued parameter, then $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \cdots \hat{\theta}_k)$ is given by the sln of simultaneous eqns

$$\frac{\partial}{\partial \theta_i} \log L = \frac{\partial}{\partial \theta_i} \log L(\theta_1, \theta_2, \cdots \theta_k) = 0 \quad \hookrightarrow (3)$$

$i = 1, 2, 3, \cdots k)$

eqns ①, ② & ③ are usually referred to as likelihood eqn for Estimating the parameters

H/w

(2) Setup an ANOVA table for the following per hectare yield in (100 kg) for the three varities of wheat grown

varities of wheat	I	II	III	IV
A1	6	7	3	8
A2	5	5	3	7
A3	5	4	3	4

$H_{01} \Rightarrow \mu_1 = \mu_2 = \mu_3$.

$H_{11} \Rightarrow \mu_1 \neq \mu_2 \neq \mu_3$

$H_{02} \Rightarrow \mu_1' = \mu_2' = \mu_3' = \mu_4'$

$H_{12} \Rightarrow \mu_1' \neq \mu_2' \neq \mu_3' \neq \mu_4'$.

varities of wheat	I	II	III	IV	T_i	T_{ij}^2
A1	-1	-2	2	-3	-4	1816
A2	0	0	2	-2	0	80
A3	0	1	2	1	4	616
$T_{.j}$	-1	-1	6	-4	G=0	32 → RSS
$T_{.j}^2$	1	1	36	16		

$N = 12$

$RSS = 32$

$CF = 0$

$TSS = 32$

$SSV = \frac{1}{3}(50) - 0 = 18$

$SST = \frac{16}{4} + \frac{16}{4}$

$SST = 8$

$SSE = TSS - SST - SSV$

$= 32 - 18 - 8$

$= 6$

Source of Variation	d.f	Sum of squares	Mean sum of squares	Variance Ratio
Treatments	3-1 = 2	SST = 8	MST = $\frac{8}{2}$ = 4	$\frac{MST}{MSE} = \frac{4}{1} = 4$
varieties	4-1 = 3	SSV = 18	MSV = $\frac{18}{3}$ = 6	
Error	2x3 = 6	SSE = 6	MSE = $\frac{6}{6}$ = 1	$\frac{MSV}{MSE} = \frac{6}{1} = 6$
total	12-1 = 11			

(2,6) (3,6)

Table value of F at 5% los at (2,6) df is 5.14 calval < Tabval

Accept H_{01}

Table value of F at 5% los at (3,6) df is 4.71 calval > Tab val Reject H_{02}

Inference:- ① There is no significant difference b/w the three varities of wheat

② There is significant difference b/w the four yields of land.

11/12/2020
Friday

Estimation :- The theory statistical inference can be divided into two major areas
① Estimation of parameters
② Testing of hypotheis

Parameter :- A statistical constant derived from the population values
Ex:- Population mean, μ
Population standaid deviation, σ

Statistic :- A statistical constant derived from the sample values is called a 'statistics'

Estimation :- ~~procedure~~ It is a procedure of Estimating a population parameter by using sample information (or) Observation

1) Point Estimation
2) Interval Estimation

Point Estimate :- A point Estimate of some population parameter θ is a single numerical value

Point Estimate :- A point Estimator is a statistic for Estimating a population parameter θ & is denoted by $\hat{\theta}$

$\mu = 25 \rightarrow$ Estimate

$\hat{\theta} \leftarrow \bar{x} = 25.2 \rightarrow$ point Estimator
↙
Sample mean

Properties of Estimators:-

A estimator is said to be a good estimator if it is

1. unbiased
2. consistant
3. efficient
4. sufficient,

(or)

1. unbiasedness
2. Consistency
3. Efficiency
4. Sufficiency

Unbiased Estimator:- A statistic $\hat{\theta}$ (say) is said to be an unbiased estimator (or) its value an ubiased estimate if & only if the mean of the sampling distribution of the Estimator equals θ i.e;

$$\text{Mean of } \hat{\theta} = E(\hat{\theta}) = \theta$$

$$E(\hat{\theta}) = \theta$$

R.V

discrete cont.

$E(x) = \lambda \exp(x)$

$E(x) = \int_{-\infty}^{\infty} x f(x)\, dx$

Sampling distribution:- sample size $= 2 = n$

sample no $= N$

$N = 5$

$n = 2$

1

2

3

4

5

if the Estimator is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the bias of the Estimate $\hat{\theta}$

when the ~~estimo~~ Estimator is unbiased,

$$E(\hat{\theta}) - \theta = 0$$

i.e., the bias is zero.

positive bias :- If $E(\hat{\theta}) > \theta$, then $\hat{\theta}$ is called positively biased

negative bias :- If $E(\hat{\theta}) < \theta$, then $\hat{\theta}$ is called negatively biased.

(1)

problem :- If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased Estimators of θ, then $w_1\hat{\theta}_1 + w_2\hat{\theta}_2$ is an unbiased Estimator of θ when $w_1 + w_2 = 1$

sln :- given that

$\hat{\theta}_1$ & $\hat{\theta}_2$ are unbiased Estimators of θ

i.e., $E(\hat{\theta}_1) = \theta \rightarrow (1)$

$E(\hat{\theta}_2) = \theta \rightarrow (2)$

Now, $E(w_1\hat{\theta}_1 + w_2\hat{\theta}_2) = w_1 E(\hat{\theta}_1) + w_2 E(\hat{\theta}_2)$

$= w_1\theta + w_2\theta$ (from eq^n (1) & (2))

$= (w_1 + w_2)(\theta)$ $(w_1 + w_2 = 1)$

$E(w_1\hat{\theta}_1 + w_2\hat{\theta}_2) = \theta$

$\therefore w_1\hat{\theta}_1 + w_2\hat{\theta}_2$ is an unbiased Estimate of θ

Theorem:- Let $x_1, x_2, \cdots x_n$ be a random sample drawn from a given population with the mean μ & variance σ^2 show that the sample mean is an unbiased estimator of population mean μ

i.e, $E(\bar{x}) = \mu$.

Proof:- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Taking Expectations on both sides

$$E(\bar{x}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n} E(x_1 + x_2 + \cdots + x_n)$$

$$= \frac{1}{n} \{E(x_1) + E(x_2) + \cdots + E(x_n)\}$$

$$= \frac{1}{n} \{\mu + \mu + \cdots + \mu\} \quad [\because E(x_i) = \mu]$$

$$= \frac{1}{n} [n\mu]$$

$$\boxed{E(\bar{x}) = \mu}$$

Theorem:- for a random sample of size n, $x_1, x_2, \cdots x_n$ taken from a finite population $s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is not an unbiased estimator of σ^2 but $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is an unbiased estimator of σ^2

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

MVUE (minimum variance unbiased estimate)
(or) most efficient unbiased Estimator

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are the two unbiased estimators of θ if and if σ_1^2 and σ_2^2 are variances of their sampling distributions and $\sigma_1^2 < \sigma_2^2$, $\hat{\theta}_1$ is said to be more unbiased estimator of θ

Note:- if $\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_k$ are unbiased estimators of the parameter θ

$\sigma_1^2, \sigma_2^2, \ldots \sigma_k^2$ are variances of there sampling distributions then $\hat{\theta}_1$ is the most efficient estimator of θ if θ $\sigma_1^2 < \sigma_2^2 < \ldots < \sigma_k^2$

e = efficaency of an estimator

$e = \frac{\text{Variance } \theta_1}{\text{Variance } \theta_i}$ where $i = 2, 3, \ldots k$ $(e < 1)$

$e = \frac{\text{var } \theta_1}{\text{var } \theta_i}$

$e = \frac{\text{var } \theta_1}{\text{var } \theta_i}$

Consistency:- An estimator $T_n = T(x_1, x_2, \ldots x_n)$ based on a random sample of size n, is (gamma) said to be a consistent estimator of $\gamma(\theta)$, $\theta \in \Theta$, $\Theta \rightarrow$ parameter space, if T_n converges to $\gamma(\theta)$

i.e., if $T_n \xrightarrow{P} \gamma(\theta)$ as $n \to \infty$

In other words T_n is a consistent estimator of $\gamma(\theta)$ if for every $\varepsilon > 0$, $\eta > 0$, there exists a positive integer $n \geq m(\varepsilon, \eta)$ such that $P\{|T_n - \gamma(\theta)| < \varepsilon\} \to 1$ as $n \to \infty \Rightarrow P\{|T_n - \gamma(\theta)| < \varepsilon\} \geq 1 - \eta \ \forall n \geq n$ where n is some very large value n.

$\gamma, \theta \to$ parameter

$T_n \to$ estimator