Unit 5: Correlation Analysis

If there exists some relationship b/w two variables then the statistical analysis of such data is called bivariate analysis.

correlation refers to the relationship of two or more variables.

correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.

The correlation expresses the relationship of two sets of variables upon eachother

Types of correlation: correlation is classified into many types,

(1) +ve & -ve correlation: +ve & -ve correlation depends upon the the direction of change of variables. If two variables tends to move together in the same direction i.e an increase in the value of 1 variable affected by the increase in the value of other variable or a decrease in the value of 1 variable is affected by a decrease in the value of another variable then the correlation is called +ve or direct correlation. If two variables tends to move together in opp directions so that an increase or decrease in the value of 1 variable is affected by a decrease or increase in the value of other variable then the correlation is called -ve or inverse correlation

(2) Simple & multiple correlation: when we study only 2 variables the relationship is described as correlation. But in multiple correlation we study more than two variables simultaneously

(3) Partial & Total Corelation:- The study of 2 variables excluding some other variables is called partial corelation.

→ In Total corelation all the facts are considered in problem

(4) Linear & non linear corelation:- If the ratio of change bw two variables is uniform then there will be a linear corelation bw them - ~~for~~ ~~and~~

→ In a non-linear or curvylinear corelation the amount of change in one variables doesnot bear a constant ratio of the amount of change in other variable

Methods of studying corelation:-

(1) Karl Pearson's coeff of corelation:- This is mathematical method for measuring the magnitude of linear relationship b/w two variables It is denoted by

$$r(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \cdot \Sigma(y-\bar{y})^2}}$$

Properties of coeff of corelation:-

(1) coeff of corelation lies b/w -1 & +1 i.e the limits for corelation are ~~-1~~ ±1

(2) If $r=1$ then the two variables are perfect +ve if $r=-1$ then the corelation is perfect -ve

If $r>0$ then there is a +ve corelation b/w two variables

If $r<0$ then there is a -ve corelation b/w two variables

If $r=0$ then two variables are uncorelated

Prob

1. A Psychological test of intelligence and engineering ability were applied to 10 students. Here is a record of data showing intelligence ratio & engineering ratio. Cal the coeff of Corelation.

intelligence ratio	105	104	102	101	100	99	98	96	93	92
eng ratio	101	103	100	98	95	96	104	92	97	94

Sol

Let x = intelligence ratio

y = eng ratio

$\bar{x} = 99$

$\bar{y} = 98$

$(x-\bar{x})$	6	5	3	2	1	0	-1	-3	-6	-7 = 0
$(y-\bar{y})$	3	5	2	0	-3	-2	6	-6	-1	-4 = 0
$(x-\bar{x})^2$	36	25	9	4	1	0	1	9	36	49 = 170
$(y-\bar{y})^2$	9	25	4	0	9	4	36	36	1	16 = 140
$(x-\bar{x})(y-\bar{y})$	18	25	6	0	-3	0	-6	18	6	28 = 92

$$r(x,y) = \frac{92}{\sqrt{23800}} = 0.5963$$

2. Cal the coeff of corelation to the following data

x :	38	45	46	38	35	38	46	32	36	38
y :	28	34	38	34	36	26	28	29	25	36

Sol

$\bar{x} = 39.2$

$\bar{y} = 31.4$

$(x-\bar{x}) = -1.2 \quad 5.8 \quad 6.8 \quad -1.2 \quad -4.2 \quad -1.2 \quad 6.8 \quad -7.2 \quad -3.2 \quad -1.2$

$(y-\bar{y}) = -3.4 \quad 2.6 \quad 6.6 \quad 2.6 \quad 4.6 \quad -5.4 \quad -3.4 \quad -2.4 \quad -6.4 \quad 4.6$

$(x-\bar{x})^2 = 1.44 \quad 33.64 \quad 46.24 \quad 1.44 \quad 17.64 \quad 1.44 \quad 46.24 \quad 51.84$

$10.24 \quad 1.44 = 211.6$

$(y-\bar{y})^2 = 11.56 \quad 6.76 \quad 43.56 \quad 6.76 \quad 21.16 \quad 29.16 \quad 11.56 \quad 5.76 \quad 40.96$

$21.16 = 198.4$

$(x-\bar{x})(y-\bar{y}) = 4.08 \quad 15.08 \quad 44.88 \quad -19.32 \quad 6.48 \quad -23.12 \quad 17.28 \quad 20.48 \quad -5.52$

$= 57.2$

$$r(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \cdot \Sigma(y-\bar{y})^2}} = \frac{57.2}{\sqrt{(211.6)(198.4)}} = 0.2792$$

Spearman's Rank corelation:- This method is based on ranks and is useful in dealing with qualitative characteristics such as intelligence honest, kindness etc. It cannot be measured quantitatively as in case of Karl Pearson's coeff of corelation. It is based on the ranks given to the observations. The formula for Spearman's Rank corelation coeff is given by

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

In case common ranks are given to the repeated values then we can use the following formula after cal the correction factor of each repeated values

$$\therefore \rho = 1 - \frac{6[\Sigma d^2 + CF]}{n(n^2-1)}$$

Properties of Rank corelation.

(1) The value of Spearman's Rank corelation is always lies b/w $-1 \leq \rho \leq +1$

(2) If $\rho = 1$ then there is a complete agreement in the order of the ranks and the directions of the ranks is same.

(3) If $\rho = -1$ then there is a complete disagreement in the order of the ranks & the direction of ranks are opposite

Prob:

1. Cal rank corelation to the following data

x:	75	95	85	80	66	52	99	72
y:	86	90	95	84	72	99	55	64

Sol:

$n = 8$

R_1:	5	2	3	4	7	8	1	6
R_2:	4	3	2	5	6	1	8	7

$d = R_1 - R_2 = 1 \quad -1 \quad 1 \quad -1 \quad 1 \quad 7 \quad -7 \quad -1$

$d^2 = 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 49 \quad 49 \quad 1 = 104$

$$\rho = 1 - \left[\frac{6\Sigma d^2}{n(n^2-1)}\right] = 1 - \left[\frac{6.24}{504}\right] = -0.2381$$

$R_1 \rightarrow$ highest ranks / ranks of x

$R_2 \rightarrow$ ranks / highest rank of y

Spearman

2. Cal rank covelation coeff of following data

x: 100 105 95 100 100 95 107 132 150 132

y: 95 100 100 96 96 95 105 96 145 136

x	y	R_1	R_2	$d = R_1 - R_2$	d^2
100	95	7	9.5	-2.5	6.25
105	100	5	2.5	0.5	0.25
95	100	9.5	2.5	5	25
100	96	7	7	0	0
100	96	7	7	0	0
95	95	9.5	9.5	0	0
107	105	4	3	1	1
132	96	2.5	7	-4.5	20.25
150	145	1	1	0	0
132	136	2.5	2	0.5	0.25
					53

In x series 132 is repeated 2 times

$m = 2$

$\therefore$ correction factor $= \frac{1}{12}(m^3 - m)$

$= 0.5$

in x series 100 is repeated 3 times

$m = 3$, $C.F = \frac{1}{12}(m^3 - m) = 2$

In x series 95 is repeated 2 times.

$m = 2, \quad CF = 0.5$

In y series 100 is repeated 2 times

$m = 2, \quad CF = 0.5$

In y series 96 is repeated 3 times

$m = 3, \quad CF = 2$

In y series 95 is repeated 2 times

$m = 2, \quad CF = 0.5 \qquad = 6$

~~0.5+6~~

Now spearman rank corelation coeff

$$\rho = 1 - \frac{6[\Sigma d^2 + C.F]}{n(n^2-1)}$$

$$= 1 - \left[\frac{6(53+6)}{10(100-1)}\right] = 0.6424$$

3. Cal S.R.C coeff to the following data

$n = 12$

x	y	R_1	R_2	$d = R_1 - R_2$	d^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	11.5	-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
69	71	4.5	4	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
					72.5

In x series 68 is repeated 2 times

$m = 2$

$CF = \frac{1}{12}(m^3 - m) = 0.5$

In x series 67 is repeated 2 times

$m = 2, \quad CF = 0.5$

In y series 68 is repeated 4 times

$m = 4, \quad CF = 5$

In y series 66 is repeated 2 times

$m = 2, \quad CF = 0.5$

In y series 65 is repeated 2 times

$m = 2, \quad CF = 0.5 = 7$

$$\rho = 1 - \frac{6[\Sigma d^2 + C.F]}{n(n^2-1)} = 1 - \left[\frac{6(72.5+7)}{12(12^2-1)}\right]$$

$$= 0.722$$

Prob

1. 10 competitors in a musical test were ranked by 3 judges A B C in the following order & using rank correlation method discuss which pair of judges has the nearest approach to common likings in music

(R_A) ranks given by judge A :	1	6	5	10	3	2	4	9	7	8
(R_B) Judge B :	3	5	8	4	7	10	2	1	6	9
(R_C) Judge C :	6	4	9	8	1	2	10	5	7	3

$d = R_A - R_B = d_1^2 =$	4	1	9	36	16	64	4	64	1	1 = 200
$d_2 = R_B - R_C = d_2^2 =$	9	1	1	16	36	64	64	16	1	36 = 244
$d_3 = R_C - R_A = d_3^2 =$	25	4	16	4	4	0	36	16	0	25 = 130

$$\rho(A,B) = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{1200}{10(10^2-1)} = -0.2121$$

$$\rho(B,C) = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{1464}{990} = -0.4788$$

$$\rho(C,A) = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{780}{990} = 0.2121$$

The rank correlation of C & A has the max value so we conclude that the judges A & C has the nearest approach to common likings in music

Regression analysis:- The study of corelation measures the direction of the relationship b/w two variables in corelation we cannot estimate the value of 1 variable when the value of other variable is given In regression we can estimate the value of 1 variable with the value of other variable which is known the statistical method which helps us to estimate the unknown value of 1 variable from the known value of related variable is called regression. regression is an avg relationship b/w two variables the line described in avg relationship b/w 2 variables is known as a line of regression.

Regression eq's:- Regression eq is an algebraic expression of the regression lines. R.eq may be classified into 2 types

① Regression eq of y on x can be written as $y-\bar{y} = byx(x-\bar{x})$

where byx is the R. coeff of y on x and is $\left(byx = r, \frac{\sigma y}{\sigma x}\right)$ defined as →

② R.eq of x on y can be written as $(x-\bar{x}) = bxy(y-\bar{y})$

where bxy is the R. coeff of x on y and is defined as $bxy = r \cdot \frac{\sigma x}{\sigma y}$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \times \Sigma(y-\bar{y})^2}}$$

$$\sigma x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \qquad \sigma y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$$

Properties of r.coeff :- The angle b/w two regression lines is given by $\tan\theta = \frac{\sigma_y}{\sigma_x}\left(\frac{1-r^2}{r}\right)$

If $r = 0$ then $\tan\theta = \infty$

$$\theta = \frac{\pi}{2}$$

If $r = \pm 1$ then $\tan\theta = 0$

$$\theta = 0$$

② The G.M of two regression coeff is equal to coeff of corelation i.e $r = \sqrt{byx \cdot bxy}$

→ If 1 r.coeff is +ve then the other R.coeff is also +ve

Prob

1. A Panel of 2 judges P & Q graded 7 performances by independently awarding marks as follows the 8th performance which judge Q could not attend was awarded 37 marks by judge P if judge Q has also been present how many marks would be expected to have been awarded by him to the 8th performance.

(x) marks by judge P :- 46 42 44 40 43 41 45

(y) Judge Q :- 40 38 36 35 39 37 41

$\bar{x} = 43$

$\bar{y} = 38$

$(x-\bar{x})$ = 3 -1 1 -3 0 -2 2

$(y-\bar{y})$ = 2 0 -2 -3 1 -1 3

$(x-\bar{x})^2$ = 9 1 1 9 0 4 4 = 28

$(y-\bar{y})^2$ = 4 0 4 9 1 1 9 = 28

$(x-\bar{x})(y-\bar{y})$ = 6 0 -2 9 0 2 6 = 21

① $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \times \Sigma(y-\bar{y})^2}} = \frac{21}{\sqrt{784}} = 0.75$

② $\sigma_x = \sqrt{\frac{28}{7}} = 2$

③ $\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \sqrt{\frac{28}{7}} = 2$

④ $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x} = 0.75 \cdot \frac{2}{2} = 0.75$

Regression eq of y on x = $y - \bar{y} = b_{yx}(x - \bar{x})$

$= y - 38 = 0.75x - 32.25$

$= y = 0.75x - 32.25 + 38$

$= y = 0.75x + 5.75$

x is given in question

$x = 37$

$y = (0.75)37 + 5.75$

$= 33.5$

2. Price index of cotton and price index of wool are given below for 12 months a year uptime the eq of line of regression b/w two indexes and also estimate the price index of wool when the price index of cotton is 100 and also estimate the price index of cotton when the price index of wool is 80

(x) P.i of cotton - 78 77 85 88 87 82 81 77 76 83 97 93

(y) P.i of wool - 84 82 82 85 89 90 88 92 83 87 78 99

$\bar{x} = 83.6667$, $\bar{y} = 88.4167$

$(x-\bar{x})$ = -5.6667 -6.6667 1.3333 4.3333 3.3333 -1.6667 -2.6667
-6.6667 -7.6667 -0.6667 13.3333 9.3333 =

$(y-\bar{y})$ = -4.4167 -6.4167 -6.4167 -3.4167 0.5833 1.5833
-0.4167 3.5833 -5.4167 0.5833 9.5833 10.5833 =

$(x-\bar{x})^2$ = 32.1115 44.4449 1.7777 18.7775 11.1109 2.779 7.1113
44.4449 58.7783 0.4445 177.7769 87.1105 = 486.6679

$(y-\bar{y})^2$ = 19.5072 41.1740 41.1740 11.6738 0.3402 2.5068
0.1736 12.8400 29.3406 0.3402 91.8396 112.0062 = 362.9962

$(x-\bar{x})(y-\bar{y})$ = 25.0281 42.7782 -8.5554 -14.8056 1.9443 -2.6389
1.1112 -23.8888 41.5282 -0.3889 127.7700 98.7771
= 288.6665

(4) $r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \times \Sigma(y-\bar{y})^2}} = \frac{6665}{\sqrt{176619.6649}} = \frac{288.68845}{} = 0.6869$

(3) $\sigma x = \sqrt{\frac{486.6679}{12}} = 6.3683$

(2) $\sigma y = \sqrt{\frac{362.9162}{\sqrt{12}}} = 5.4994$

(4) $byx = r \cdot \frac{\sigma y}{\sigma x} = 0.6869 \cdot \frac{5.4994}{6.3683} = 0.5932$

(5) $bxy = r \cdot \frac{\sigma x}{\sigma y} = 0.6869 \times \frac{6.3683}{5.4994} = 0.7954$

Regression eq of y on x $= y - \bar{y} = byx(x - \bar{x})$

$= y - 88.4167 = 0.5932(x - 83.6667)$

$= y = 0.5932x - 49.6311 + 88.4167$

$= y = 0.5932x + 38.7856$

$= x = 100$

$y = 0.5932(100) + 38.7856$

$= 98.1056$

x & y are in the question

Regression eq of x on y $= x - \bar{x} = bxy(y - \bar{y})$

$= x - 83.6667 = 0.7954(y - 88.4167)$

$= x = 0.7954y - 70.3266 + 83.6667$

$= x = 0.7954y + 13.3401$

$= y = 80$

$x = 76.9721$

3. The eq of 2 regression lines are $7x - 16y + 9 = 0$, $5y - 4x - 3 = 0$ find the means of x & y and also find coeff of correlation

Sol

W.K.T the 2 regression lines passes through the avg of x & y i.e $7\bar{x} - 16\bar{y} + 9 = 0$, $5\bar{y} - 4\bar{x} - 3 = 0$

~~$7\bar{x} - 16\bar{y} + 9 = 0$~~

$28\bar{x} - 64\bar{y} + 36 = 0$

$28\bar{x} - 35\bar{y} + 21 = 0$

$-29\bar{y} + 15 = 0$

$\bar{y} = \frac{15}{29}$

$\bar{y} = 0.5172$

$5\bar{x} = 7\bar{x} - 16(0.5172) + 9 = 0$

$\bar{x} = 0.103$

Let $7x - 16y + 9 = 0$ be the eq of x on y

$7x = 16y - 9$

$x = \frac{16}{7}y - \frac{9}{7}$

which is in the form of $x = b_{xy} \cdot y + a$

$b_{xy} = \frac{16}{17}$

Let $5y - 4x - 3 = 0$ be the eq of y on x

$5y = 4x + 3$

$y = \frac{4x}{5} + \frac{3}{5}$

which is in the form of $y = b_{yx} \cdot x + a$

$b_{yx} = \frac{4}{5}$

$r = \sqrt{byx \, bxy}$

$= \sqrt{\frac{16}{7} \times \frac{4}{5}} = 1.3522$

$r = 1.3522$ which is incorrect because r should be lies b/w -1 & $+1$ so consider $7x - 16y + 9 = 0$ as eq of y on x &

$5y - 4x - 3 = 0$ as eq of x on y

$7x - 16y + 9 = 0$

$-16y = -7x - 9$

$y = \frac{7}{16}x + \frac{9}{16}$

$5y - 4x - 3 = 0$

$-4x = -5y + 3$

$x = \frac{5}{4}y - \frac{3}{4}$

$byx = \frac{7}{16}$

$bxy = \frac{5}{4}$

$r = \sqrt{byx \, bxy} = 0.7395$

4. The following data based on 450 students are given for marks in statistics & economics at a certain examination the mean marks in statistics is 40. The mean marks in economics is 80 the S.D of marks in statistics is 12 variance of marks in economics is 256 & sum of products deviations of marks from their respective means is 42075 estimate the avg marks in economics who obtain 50 marks in statistics.

Sol: Let x = marks in statistics

y = marks in economics

$N = 450$

$\bar{x} = 40, \ \bar{y} = 80, \ \sigma x = 12, \ \sigma y = 16$

$\sigma^2 = 256 = 16$

$\Sigma(x-\bar{x})(y-\bar{y}) = 42075$

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{N \times \sigma_x \sigma_y} = \frac{42075}{450 \times 12 \times 16} = 0.4870$$

$$b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x} = 0.4870 \times \frac{16}{12} = 0.6493$$

The regression eq^n of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 80 = 0.6493(x - 40)$$

$$y - 80 = 0.6493x - 25.9720$$

$$y = 0.6493x + 54.0280$$

$$x = 5.0$$

$$y = 86.4928$$

5. The height of mothers & daughters are given in the following table estimate the avg height of daughter when the height of mother is 64.5 inches.

(x) height of mom (in inches)	62	63	64	64	65	66	69	70
(y) daughter	64	65	61	69	68	71	69	69
							65	

$\bar{x} = 65.25$ $\quad \bar{y} = 66.25$

$(x-\bar{x})$ = -3.25 -2.25 -1.25 -1.25 -0.25 0.75 2.75 4.75

$(y-\bar{y})$ = -2.25 -1.25 -5.25 0.75 1.75 4.75 -1.25 2.75

$(x-\bar{x})^2$ = 10.5625 5.0625 1.5625 1.5625 0.0625 0.5625 7.5625
22.5625 = 49.5

$(y-\bar{y})^2$ = 5.0625 1.5625 27.5625 0.5625 3.0625 22.5625 1.5625
7.5625 = 69.5

$(x-\bar{x})(y-\bar{y})$ = 7.3125 2.8125 6.5625 -0.9375 -0.4375 3.5625 -3.4375
13.0625 = 28.5

① $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \times \Sigma(y-\bar{y})^2}} = \dfrac{28.5}{\sqrt{3440.25}} = 0.4859$

② $\sigma_x = \sqrt{\dfrac{49.5}{8}} = 2.4875$

③ $\sigma_y = \sqrt{\dfrac{69.5}{8}} = 2.9475$

④ $b_{yx} = r \cdot \dfrac{\sigma_y}{\sigma_x} = 0.4859 \cdot \dfrac{2.9475}{2.4875} = 0.5758$

Regression eq. of y on x = $y - \bar{y} = b_{yx}(x - \bar{x})$

$= y - 66.25 = 0.5758(x - 65.25)$

$= y - 66.25 = 0.5758x - 37.5680$

$x = 64.5$

$y - 66.25 = 37.1391 - 37.5680$

$y = 65.8211$

Analysis of Variance - (ANOVA)

ANOVA is a statistical tool to test the homogenity of diff groups based on their differences. ANOVA is is used to determine the diff blw the means of samples by analysing the variation within each of the samples. ANOVA may be classified into two ways.

(1) one way classification (O.W.C)

(2) Two way classification (T.W.C)

(1) O.W.C is a shortcut method where a single factor is considered and its effect on samples is observed this method is performed when the mean of the samples are equal or not. For eg: To study the effect of diff levels of fertilisers on a plant growth

(2) To study the effect of diff levels of a medicine on a disease.

consider a single factor and its effect on k samples as

1	2	3	----	k
x_{11}	x_{21}	x_{31}	----	x_{k1}
x_{12}	x_{22}	x_{32}	----	x_{k2}
⋮	⋮	⋮		⋮
⋮	⋮	⋮		⋮
x_{1n}	x_{2n}	x_{3n}		x_{kn}

To test there is any significal diff b/w samples or treatments means, let us consider the null hypothesis as

$H_0: \mu_1 = \mu_2 = \cdots = \mu_t$

$H_1: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_t$

To test the above null hypothesis we can use F-ratio test as follows

① $\sum\sum x_{ij}^2$

② Find grand total G

* Cal correction factor as $CF = \frac{G^2}{n}$

* Cal total sum of squares $TSS = \sum\sum x_{ij}^2 - CF$

* Cal sum of squares for treatments as $SST = \frac{\sum x_i^2}{n_i} - CF$

* Find the sum of squares of error as (SSE)

$SSE = TSS - SST$

* Now cal the mean sum of squares for treatment as MSST

$$MSST = \frac{SST}{k-1}$$ k = no. of treatments or samples

* & mean sum of square error is

$$MSSE = \frac{SSE}{n-k}$$

Here n = total no. of obs, k = no. of samples (or) treatments

k-1 is treatment degrees of freedom

n-k is error degrees of freedom

Now Cal F-ratio as $F = \frac{MSSt}{MSSE}$

Now compare the cal value of F at 5% or 1% level of significance with (k-1, n-k) degrees of freedom if the table value of F is greater than the cal value of F then we accept our null hypothesis otherwise we reject the hypothesis.

Prob

1. Following are the data results of samples tested for medication, excercises and diet from 3 population. test the hypothesis that there is any diff blw there 3 areas at 1% level of confidence

Medication	10	12	9	15	13	= 59
Exercise	6	8	3	0	2	= 19
diet	5	9	12	8	4	= 38
						116

Sol $k = 3$; $n = 15$

H_0 = There is no significant diff among medication, excercise & diet

H_1 = There is a significant diff among medication, excercise & diet

$\sum\sum x_i^2 = 10^2 + 12^2 + 9^2 + 15^2 + 13^2 + 6^2 + 8^2 + 3^2 + 2^2 + 5^2 + 9^2 + 12^2 + 8^2 + 4^2 = 1162$

$G = 116$

$CF = \frac{G^2}{n} = \frac{(116)^2}{15} = 897.0667$

$TSS = \sum\sum x_{ij}^2 - CF = 1162 - 897.0667$

$= 264.9333$

$SSt = \frac{\sum x_i^2}{n_i} - CF$

$= \frac{(59)^2}{5} + \frac{(19)^2}{5} + \frac{(38)^2}{5} - 897.0667 = 160.1333$

$SSE = TSS - SSt$

$= 264.9333 - 160.1333$

$= 104.8$

$MSSt = \frac{SSt}{k-1} = \frac{160.1333}{2} = 80.0667$

$MSSE = \frac{SSE}{n-k} = \frac{104.8}{12} = 8.7333$

$F = \frac{MSSt}{MSSE} = \frac{80.0667}{8.7333} = 9.1680$

The table value of F with $(k-1, n-k) = (2, 12)$ at 1% level of Significance is 6.93

Hence the table value of F is less than the cal value of F so we reject our null hypothesis

(2) Two way classification: T.w.c technique is used when the data are classified on the basis of two factors.

for eg: The agricultural O/P maybe classified on the basis of diff varieties of seeds and also on the basis of diff varieties of fertilizers. Let us consider the first factor has 'k' samples and second factor has 'n' samples are listed in a table below

	C_1	C_2	----	C_k
R_1	x_{11}	x_{12}	---	x_{1k}
R_2	x_{21}	x_{22}	---	x_{2k}
	⋮			
R_n	x_{n1}	x_{n2}	---	x_{nk}

Let R = no. of rows

C = no. of columns

n = total no. of obs

To test the significance diff. of effects of two factors

Let us consider H_0 as there is no significant diff. of effect of first factors & there is no significant diff of effect of second factor on the given samples

First find

(1) $\sum\sum x_{ij}^2$ and then find grand total G

(2) Find the correction factor $CF = \frac{G^2}{n}$

(3) Find total sum of squares $TSS = \sum x_{ij}^2 - CF$

(4) Find the rows sum of squares $SSR = \frac{\sum R_i^2}{n_i} - CF$

(5) Find column sum of squares $SSC = \frac{\sum C_j^2}{n_j} - CF$

(6) Find error sum of squares $SSE = TSS - SSR - SSC$

(7) Find mean sum of squares of the rows as $MSSR = \frac{SSR}{R-1}$

(8) Find " " " " " columns as $MSSC = \frac{SSC}{C-1}$

(9) Find " " " " " error as $MSSE = \frac{SSE}{n-R-C+1}$

(10) Now cal F-Ratio for the first factor (row wise) is $\frac{MSSR}{MSSE}$

(11) cal F ratio for the 2nd factor (column wise) is $\frac{MSSC}{MSSE}$

If cal value of F is less than the table value of F with $(R-1, N-R-C+1)$ degrees of freedom at 5% or 1% level of significance then we accept our first factor null hypothesis

If cal values of F is less than the table value of F with $(C-1, n-R-C+1)$ degrees of freedom at 5% or 1% level of significance the we accept our 2nd factor null hypothesis

Prob

1. Use 2-way classification test the significance of students of subs

	A	B	C	
Bio	157	250	156	= 563
French	148	140	158	= 446
English	150	150	145	= 445

Maths	140	140	151	= 431
Science	166	185	166	= 511
Computer	167	200	161	= 528
	922	1025	937	2924

Let rows represent subjects and column represent student

R = no. of rows = 6

C = no. of columns = 3

n = total no. of obs = 18

H_0: There is no significant diff among the subs

H_{0_2}: There is no " " among the students

H_1: There is significant diff among the subs

H_{12}: " " " " " " students

$\sum\sum x_{ij}^2 = 487170$

$G = 2924$

$CF = \frac{G^2}{n} = \frac{2924^2}{18} = 474987.5556$

$TSS = \sum\sum x_{ij}^2 - CF = 12182.4444$

$SSR = \frac{\sum R_i^2}{M_i} - CF$

$= \frac{(563)^2}{3} + \frac{(446)^2}{3} + \frac{(445)^2}{3} + \frac{(431)^2}{3} + \frac{(511)^2}{3} + \frac{(528)^2}{3} - CF$

$= 4871.1111$

$SSC = \frac{\sum C_i^2}{n_i} - CF$

$= \frac{(922)^2}{6} + \frac{(1025)^2}{6} + \frac{(937)^2}{6} - CF = 2088.7777$

$SSE = TSS - SSR - SSC = 5252.5556$

$MSSR = \frac{SSR}{R-1} = 974.2222$

$MSSC = \frac{SSC}{C-1} = 1029.3889$

$MSSE = \frac{SSE}{n-R-C+1} = 525.2556$

$F_R = \frac{MSSR}{MSSE} = 1.8548$

$F_C = \frac{MSSC}{MSSE} = 1.9598$

The table value of F_R with $(R-1, n-R-C+1) = (5, 10)$ at 5% level of significance is 3.33.

∴ The cal value of F_R is less than table of F_R so we accept our null hypothesis (1)

The table value of F_C with $(C-1, n-R-C+1) = (2, 10)$ degrees of freedom at 5% level of significance is 4.10

∴ The cal value of F_C is less than the table value of F_C so, we accept our null hypothesis (2)

The following data consists of breaking strength of 8 widgets were made from each of 4 materials and the resulting strength are shown below Test the significant reference on the widgets of materials & breaking strength of widgets.

widgets	A	B	C	D	
1.	59	27	13	15	= 94
2.	57	43	32	28	= 160
3.	42	25	19	11	= 97
4.	32	11	11	12	= 66
5.	43	39	30	34	= 146
6.	50	49	30	34	= 163
7.	31	27	16	13	= 87
8.	51	34	28	23	= 136
	345	255	179	170	949

Let rows represents the widgets and columns represents the material

r = no of rows = 8

c = no of col = 4

$n = 32$ (1)

H_{01} = There is no significant diff on the breaking strength of widgets

H_{02} = There is no significant diff on the strength of material

H_{11} = There is significant diff on the breaking strength of widgets

H_{12} = " " " " " " strength of material

$\sum\sum x_{ij}^2 = 33399$

$G = 949$

$CF = \frac{G^2}{n} = 28143.7813$

$TSS = \sum\sum x_{ij}^2 - CF = 5255.2187$

$SSR = \left(\frac{R_1^2}{4} + \frac{R_2^2}{4} \ldots\right) - CF = 2393.9687$

$SSC = \left(\frac{C_1^2}{8} + \frac{C_2^2}{8} \ldots\right) - CF = 2480.0937$

$SSE = TSS - SSR - SSC = 381.1563$

$MSSR = \frac{SSR}{R-1} = \frac{2393.9687}{7} = 341.9955$

$MSSC = \frac{SSC}{C-1} = \frac{2480.0937}{3} = 826.6979$

$MSSE = \frac{SSE}{n-R-C+1} = \frac{381.1563}{(21)} = 18.1503$

$F_R = \frac{MSSR}{MSSE} = 18.8424$ $\quad F_C = \frac{MSSC}{MSSE} = 45.5473$

The table value of F_R with $(R-1, n-R-C+1) = (7, 21)$ at 5% is 2.49

The cal value of F is > table value of F so we reject our null hypothesis

The table value of F_c with $(C-1, n-k-C+1)$ $(3, 21)$ at 5% 3.07 the cal value of F is > than the table value of F so we reject our null hypothesis