CHAPTER-V

CORRELATION ANALYSIS

So far we have studied descriptive and inferential techniques that involve a single quantitative variable. In this chapter, we shall study the inferential techniques that involve two variables that are studied simultaneously for each individual. If for every value of a variable X we have a corresponding value of another variable Y then the series of pairs of values (X, Y) is known as a bivariate distribution. For example, a series of pairs of the ages of husbands and their wives at the time of marriage form a bivariate distribution.

In a bivariate distribution if the change in one variable appears to be accompanied by a change in the other variable and vice-versa then the variables are said to be correlated. The term correlation is mutual relationship between two or more variables. If for an increase (decrease) in the variable X, variable Y also shows increase (decrease) accordingly, the correlation will be positive. If the increase (decrease) in X, is accompanied by a corresponding decrease (increase) in variable Y, then the correlation will be negative. If change in one variable does not show a change in the other variable, the two variables are called uncorrelated or independent.

5.1 Methods of Measuring Correlation:

Scatter diagram, Karl Pearson coefficient of correlation and Spearmanôs rank correlation coefficient are the frequently used methods in correlation studies.

5.1.1 Scatter Diagram Method:

Let (X_i, Y_i), i =1, 2,·····,n be the n pairs of observations. If these paired observations are plotted on a graph paper in the XY-plane such that each pair is represented by dot in the diagram. The diagram so obtained is known as scatter or dot diagram. By looking at the scatter diagram of the various points, we can form an idea as to whether the variables are correlated or not. If all the points lie on a straight line arising from the lower left hand corner to the upper right hand corner, correlation is said to be perfect positive i.e. r = + 1. On the other hand, if all the points lying on a straight line falling from the upper left hand corner to the lower right hand corner, then correlation is said to be perfectly negative. If the plotted points fall in a narrow band, there would be a high degree of correlation

between the variables. Correlation shall be positive if points show a rising tendency from lower left hand corner to the upper right hand corner and negative if the points show a declining tendency from the upper left hand corner to the lower right hand corner of the diagram. On the other hand, if the points are widely scattered over the diagram, then we expect no correlation or poor correlation between X and Y.

Limitations of Scatter Diagram:

Scatter diagram only tells about the nature of the relationship whether it is positive or negative and whether it is high or low. It does not provide the extent of the measure of relationship between the variables. It is subjective in nature and different individuals may have different interpretations from the same set of data.

5.1.2 Karl Pearson Coefficient of Correlation or Product Moment Correlation Coefficient or Simple Correlation Coefficient:

It is a mathematical method for measuring the degree or strength of linear relationship between two variables and was suggested by Karl Pearson (1901). The Pearson coefficient of correlation between X and Y is denoted by the symbol r_{xy} or simply ‘r’ and is defined as the ratio of covariance between X and Y to the product of the standard deviations of X and Y.

$$r = \frac{Cov(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2}\sqrt{\sum (Y_i - \bar{Y})^2}}$$

where σ_x and σ_Y are the standard deviations of X and Y. Covariance and standard deviations can be written in terms of sum of squares and sum of products also.

$$Cov(X, Y) = \sum (X_i - \bar{X})(Y_i - \bar{Y}) / n$$

$$\sigma_X = \sqrt{\sum (X_i - \bar{X})^2 / n} \text{ and } \sigma_Y = \sqrt{\sum (Y_i - \bar{Y})^2 / n}$$

Here $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y and n is the number of paired observations. A simplified formula for computation of correlation coefficient is:

$$r = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sqrt{[\sum X_i^2 - (\sum X_i)^2/n][\sum Y_i^2 - (\sum Y_i)^2/n]}}$$

Shortcut and Step Deviation Method:

Since coefficient of correlation is independent of change of origin and change of scale, so the step deviation or short cut method can be used.

Case-I: If both change of origin and change of scale are considered

$$\text{i.e. } U = \frac{X - a}{h} \text{ and } V = \frac{Y - b}{k}$$

$$\text{then } r = \frac{\sum U_i V_i - (\sum U_i)(\sum V_i)/n}{\sqrt{\left[\sum U_i^2 - (\sum U_i)^2/n\right]\left[\sum V_i^2 - (\sum V_i)^2/n\right]}}$$

Case-II: If only change of origin is considered

i.e. dx= X-a and dy = Y-b

then

$$r = \frac{\sum dxdy - (\sum dx)(\sum dy)/n}{\sqrt{\sum dx^2 - (\sum dx)^2/n}\sqrt{\sum dy^2 - (\sum dy)^2/n}}$$

Properties:

i) It is a unit less number i.e. independent of the units of measurement.

ii) The correlation coefficient always lies between -1 & +1 i.e. -1 Ò r Ò 1.

iii) The coefficient of correlation is independent of change of scale and shift of origin of the variables X and Y.

iv) If two variables are independent, their correlation coefficient is zero but the converse is not true. It is because correlation coefficient measures only linear type of relationship, thus even if it is zero, the variables may have a non-linear relationship.

v) The degree of relationship between two variables is symmetrical i.e. $r_{yx} = r_{xy}$

Correlation for a Bivariate Frequency Distribution:

If the number of observations is very large and are divided into classes, in such situations the pairs of observations are represented in the form of a bivariate frequency distribution. For a bivariate frequency distribution of X and Y the correlation coefficient is calculated by the following formula:

$$r_{xy} = \frac{\sum f_{xy} XY - (\sum f_x X)(\sum f_y Y)/N}{\sqrt{[\sum f_x X^2 . (\sum f_x X)^2/n][\sum f_y Y^2 . (\sum f_y Y)^2/n]}}$$

where $N = \Sigma f_x = \Sigma f_y = \Sigma\Sigma f_{xy}$ and X and Y are mid values.

Testing the Significance of Population Correlation Coefficient

Case-1: Testing of ρ = 0 i.e. to test whether the variables in the population are linearly uncorrelated.

Step-wise Procedure:

i) $H_0 : \rho = 0$

ii) $H_1 : \rho \neq 0$ (Two Tailed Test)

$H_1 : \rho > 0$ (Right Tailed Test)

$H_1 : \rho < 0$ (Left Tailed Test)

iii) Select level of significance α

iv) Test statistic under H_0

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ follows student's t-distribution with (n-2) d.f.

v) Decision:

t_{cal} is compared with its critical t-value at (n-2) d.f. and level of significance α

Two Tailed Test Reject H_0 if $|t_{cal}| > t_{\alpha/2(n-2)}$

Right Tailed Test Reject H_0 if $t_{cal} > t_{\alpha(n-2)}$

Left Tailed Test Reject H_0 if $t_{cal} < -t_{\alpha(n-2)}$

Example-1: A random sample of married couples shows the age of husbands (x) and their wives (y) in different years as:

x:	30	29	36	72	37	36	51	48	37	50	51	36
y:	27	20	34	67	35	37	50	46	36	42	46	35

Calculate the coefficient of simple correlation between age of husbands and their wives and test for its significance

Solution: n = 12, $\Sigma x = 513$, $\Sigma y = 475$

$\Sigma x^2 = 23557$, $\Sigma y^2 = 20385$ and $\Sigma xy = 21861$

$$r = \frac{\Sigma xy - (\Sigma x)(\Sigma y)/n}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}} = \frac{21861 - \frac{513 \times 475}{12}}{\sqrt{\left[23557 - \frac{513^2}{12}\right]\left[20385 - \frac{475^2}{12}\right]}}$$

$$= \frac{21861 - 20306}{\sqrt{[23557 - 21930.75][20385 - 18802.08]}}$$

$$= \frac{1555}{\sqrt{1626.25 \times 1586.92}} = \frac{1555}{1606.44} = 0.969$$

H_0: $\rho = 0$ (age of husbands and their wives are independent)

H_1: $\rho \neq 0$ (Two tailed test)

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.969 \times \sqrt{12-2}}{\sqrt{1-0.969^2}} = \frac{0.969 \times \sqrt{10}}{\sqrt{1-0.939}} = 12.4$$

Since $|t_{cal}|$ > table t-value (2.228) at 10 df and $\alpha = 0.05$, therefore, we conclude that age of husband has a high positive correlation with the age of wife.

Example-2: Compute the Karl Pearson correlation coefficient from the following data and test for its significance.

Solution: **X : 9 8 7 6 5 4 3 2 1**

Y : 15 16 14 13 11 12 10 8 9

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

$\overline{X}$ = × X/n = 45/9 = 5

$\overline{Y}$ = × Y/n = 108/9 = 12

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

Test of Significance: $H_0 : ɟ = 0$ vs $H1 : ɟ Í 0$ (Two tailed test)

Let Ŭ= 0.05

Test statistic: $t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.95\sqrt{9-2}}{\sqrt{1-(0.95)^2}} = \frac{0.95\sqrt{7}}{\sqrt{1-0.9025}} = \frac{2.513}{\sqrt{.0975}} = 8.05$

Since $|t_{cal}| = 8.05 > t_{tab} = 2.36$ at 7 d.f., therefore we reject H_0 and conclude that there is significant correlation among the variables in the population.

Example-3: Calculate the coefficient of correlation between total cultivable area (X) and the area under wheat (Y) from the following bivariate distribution of data selected from 66 villages.

Area under Wheat (in hectare)	**Total cultivable Area (in hectare)**					
	0-200	**200-400**	**400-600**	**600-800**	**800-100**	**Total**
0-50	12	6	-	-	-	18
50-100	2	18	4	2	1	27
100-150	-	4	7	3	-	14
150-200	-	1	-	2	1	4
200-250	-	-	-	1	2	3
Total	**14**	**29**	**11**	**8**	**4**	**66**

Solution:

Mid Values	**X**	**100**	**300**	**500**	**700**	**900**				
Y	**u** **v**	**-2**	**-1**	**0**	**1**	**2**	**Total f_y**	**f_yv**	**f_yv^2**	**$f_{xy}uv$**
25	-2	12 (48)	6 (12)				18	-36	72	60
75	-1	2 (4)	18 (18)	4 (0)	2 (-2)	1 (-2)	27	-27	27	18
125	0	-	4 (0)	7 (0)	3 (0)	-	14	0	0	0
175	1	-	1 (-1)	-	2 (2)	1 (2)	4	4	4	3
200	2	-	-	-	1 (2)	2 (8)	3	6	12	10
	Total f_x	**14**	**29**	**11**	**8**	**4**	**66**	**-53**	**115**	**91**
	f_xu	**-28**	**-29**	**0**	**8**	**8**	**-41**			
	f_xu^2	**56**	**29**	**0**	**8**	**16**	**109**			
	f_xuv	**52**	**29**	**0**	**2**	**8**	**91**			

From the table we get

$$N = 66,\ \Sigma f_x u = -41,\ \Sigma f_x u^2 = 109,\ \Sigma f_y v = -53,\ \Sigma f_y v^2 = 115 \text{ and } \Sigma f_{xy}\, uv = 91$$

$$\text{Thus } r_{xy} = \frac{\Sigma f_{xy} uv - (\Sigma f_x u)(\Sigma f_y v)/N}{\sqrt{\left[\Sigma f_x u^2 - (\Sigma f_x u)^2/N\right]\left[\Sigma f_y v^2 - (\Sigma f_y v)^2/N\right]}}$$

$$= \frac{91 - (-41)(-53)/66}{\sqrt{\{109 - (-41)^2/66\}\{115 - (-53)^2/66\}}} = 0.749$$

Thus the correlation coefficient between total cultivable area and the area under wheat is +0.749.

Example-4: For a random sample of 18 paired observations from a bivariate normal population, the correlation coefficient is obtained as -0.6. Test the significance of correlation in the population at Ŭ= 0.05.

Solution: Set up the null and alternative hypothesis as:

i) $H_0 : ɟ = 0$

ii) $H_1 : \rho \neq 0$ (Two Tailed Test)

iii) $\alpha = 0.05$

iv) Compute the test statistic

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.6\sqrt{18-2}}{\sqrt{1-0.36}} = \frac{-0.6(4)}{\sqrt{0.64}}$$

$$= \frac{-2.4}{0.8} = -3.00 \text{ with } (18-2) = 16 \text{ d.f.}$$

$|t_{cal}| = 3.00$

t_{tab} for two tailed test at $\alpha = 0.05$ $(t_{0.025,\ 16}) = 2.12$

Since $|t_{cal}| > t_{tab}$, therefore, we reject H_0 and conclude that there is significant correlation between the variables in the population.

Case-II: Testing of $\rho = \rho_0$ $(\rho_0 \neq 0)$

For testing the significance of correlation coefficient for a bivariate normal population in which $\rho \neq 0$, i.e. in non-null case, Prof. R.A. Fisher proved that the sampling distribution of r is by no means student's t and suggested the use of Fisher Z-transformation.

$Z_r = \frac{1}{2}\log_e \frac{1+r}{1-r}$ and proved that even for small samples, the distribution of Z_r is approximately normal with mean $Z_\rho = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho}$ and variance 1/(n-3) and for large values of n (say n>50), the approximation is fairly good.

Step-wise Procedure:

1. $H_0 : \rho = \rho_0$ $(\rho_0 \neq 0)$
2. $H_1 : \rho \neq \rho_0$
3. Select α
4. Test statistic under H_0

$$Z_r = \frac{1}{2}\log_e \frac{1+r}{1-r} = 1.1513\log_{10} \frac{1+r}{1-r}$$

$$Z_\rho = \frac{1}{2}\log_e \frac{1+\rho_0}{1-\rho_0} = 1.1513 \log_{10} \frac{1+\rho_0}{1-\rho_0}$$

Thus $Z_r \sim N\left(Z_\rho, \frac{1}{n-3}\right)$

Or $Z = \frac{Z_r - Z_p}{\sqrt{\frac{1}{n-3}}} = (Z_r - Z_\rho)\sqrt{n-3}$

If $|Z_{cal}| \geq z_{tab}$, at a specified α, then we reject H_0 and conclude that correlation coefficient in the population is significantly different from ρ_0.

Example-5: A correlation coefficient of 0.5 is obtained from a sample of 19 pairs of observations. Can the sample be regarded as drawn from a bivariate normal population in which true correlation coefficient is 0.7?

Solution:

i) $H_0 : \rho = 0.7$ and

ii) $H_1: \rho \neq 0.7$ (Two tailed test)

iii) Choose level of significant $\alpha = 0.05$

Applying Fisher Z-transformation, we get:

$$Z_r = 1.1513 \log_{10} \frac{1+r}{1-r} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5}$$

$= 1.1513 \log_{10} 3 = 1.1513\ (0.4771) = 0.5492$

and Mean $Z_P = 1.1513 \log_{10} \frac{1+\rho}{1-\rho} = 1.1513 \log_{10} \frac{1+0.7}{1-0.7}$

$= 1.1513 \log_{10} 5.67 = 1.1513\ (0.7536) = 0.8676$

Computing Z-test statistic, we get:

$$Z_{cal} = \frac{Z_r - Z_p}{1/\sqrt{n-3}} = (Z_r - Z_p)\sqrt{n-3}$$

$= (0.5492 - 0.8676)\sqrt{16} = -0.3184(4) = -1.27$

Since the $|Z_{cal}| = 1.27$ is less than the tabulated value $Z_{\alpha/2} = 1.96$ at 5% level of significance, therefore, we do not reject H_0 and conclude that the sample may be regarded as coming from a bivariate normal population with $\rho = 0.7$.

Testing the Significance of Difference between two Independent Correlation Coefficients:

The above case concerning single correlation coefficient can be generalized to test the significance of difference between two independent correlation coefficients. Let r_1 and r_2 be the sample correlation coefficients observed in two independent samples of size n_1 and n_2, respectively. Then

$$Z_1 = \log_e\left(\frac{1+r_1}{1-r_1}\right) \text{ and } Z_2 = \log_e\left(\frac{1+r_2}{1-r_2}\right)$$

Testing Procedure:

i) $H_0 : \rho_1 = \rho_2 = \rho$ (say) Correlation coefficients do not differ significantly i.e. the samples are drawn from the same bivariate normal population or from the different populations with the same correlation coefficient ρ.

ii) $H_1: \rho_1 \neq \rho_2$ (Two Tailed Test)

iii) Choose level of significance α

iv) Test Statistic (Under H_0)

$$Z = \frac{Z_1 - Z_2}{\sqrt{V(Z_1 - Z_2)}} = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0,1)$$

(v) Conclusion

If $|Z_{cal}| \geq z_{\alpha/2}$, then reject H_0.

Example-6: Two independent samples of 23 and 28 pairs of observations were analyzed and their correlation coefficients were found as 0.5 and 0.8, respectively. Do these values differ significantly?

Solution:

i) $H_0 : \rho_1 = \rho_2$, i.e. correlation coefficients do not differ significantly i.e. the samples are drawn from the same population.

ii) $H_1 : \rho_1 \neq \rho_2$ (Two tailed test)

iii) $\alpha = 0.05$

iv) Test Statistic:

$$Z_1 = 1.1513 \log_{10} \frac{1+r_1}{1-r_1} = 1.1513 \log_{10} \frac{1+0.5}{1-0.5}$$

$= 1.1513 \log_{10} 3 = 0.55$

$$Z_2 = 1.1513 \log_{10} \frac{1+r_2}{1-r_2} = 1.1513 \log_{10} \frac{1+0.8}{1-0.8}$$

$= 1.1513 \log_{10} 9 = 1.10$

$$Z_{cal} = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} = \frac{0.55 - 1.10}{\sqrt{\frac{1}{20} + \frac{1}{25}}} = \frac{-0.55}{0.30} = -1.83$$

Since $|Z_{cal}| = 1.83$ is less than $Z_{tab} = 1.96$ at 5% level of significance (two tailed test), therefore we do not reject H_0. Thus we conclude that the correlation values do not differ significantly.

5.1.3 Rank Correlation Coefficient:

Sometimes we come across statistical series in which the variables are not capable of quantitative measurement but can be arranged in the serial order. This happens when we are dealing with qualitative data. In such cases, Charleôs Edward Spearman, a British psychologist developed a formula in 1904 which gives the correlation coefficient between the ranks of n individuals on the two attributes under study.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i denotes the difference between the ranks of i^{th} paired observation i = 1, 2, é .., n

Properties of Spearman's Rank Correlation:

i) Spearmanôs formula is the only formula to be used for finding the correlation coefficient if we are dealing with qualitative data.

ii) Spearmanôs correlation can be applied even when the data do not follow normal distribution. It is the distribution free measure as it does not make any assumption about the population from which the samples are drawn.

iii) Spearmanôs formula and Karl Pearson formula give the same value if they are applied on the same data, provided no item is repeated or all items of the series are different.

iv) The limits of Spearmanôs rank correlation coefficient are from -1 to +1.

Note: The Spearmanôs formula cannot be applied in case of bivariate frequency distribution.

Equal Ranks/Repeated Ranks: When equal ranks are assigned to some entries, an adjustment in the above formula for calculating the coefficient of rank correlation is made. The adjustment consists of adding $(m^3-m)/12$ in the value of $\times d_i^2$, where m stands for the number of individuals whose ranks are common. If there are more than one such groups of individuals with common ranks, then this value is added as many times as the number of such groups and formula can be written as:

$$r_s = 1 - \frac{6[\sum d_i^2 + (m^3 - m)/12 + ...]}{n(n^2 - 1)}$$

5.1.4 Partial and Multiple Correlations: Quite often there is inter relation between various variables recorded during any study and consequently value of one variable is simultaneously influenced by many other variables. When more than two variables are involved in a study, four major problems may arise:

i) We may be interested in studying the inter-dependence or correlation of two variables only when other variables included in study are kept constant. This is the problem of partial correlation.

ii) We may also be interested in studying the correlation between dependent variable and a number of independent variables. This is the problem of multiple correlation.

iii) We may wish to examine the effect of number of independent variables upon the dependent variable. This is the problem of multiple regression.

iv) We may be interested in studying the effect of one independent variable upon the dependent variable when the effect of all other independent variables are eliminated and this is the problem of partial regression.

Partial Correlation: The partial correlation may be defined as a statistical tool measuring the linear relationship between two variables when all other variables involved in the study are kept constant or when their linear effects are eliminated. We denote $r_{12.3}$

as the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant and is computed as under:

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

Partial correlation coefficients such as $r_{12.3}$, $r_{13.2}$ are often referred to as first order partial correlation coefficients since one variable has been held constant. Further, $r_{12.34}$, $r_{13.24}$ and $r_{14.23}$ etc. are called second order partial correlation coefficients since two variables are kept constant and these may be computed as:

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} \cdot r_{24.3}}{\sqrt{1 - r^2_{14.3}} \cdot \sqrt{1 - r^2_{24.3}}}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2} \cdot r_{34.2}}{\sqrt{1 - r^2_{14.3}} \sqrt{1 - r^2_{34.2}}}$$

First order partial correlations are tested using ά ô test

$$t = \frac{r_{ab.c} \sqrt{(n-3)}}{\sqrt{1 - r^2_{ab.c}}} \text{ with (n-3) degrees of freedom}$$

where, $r_{ab.c}$ is the first order partial correlation between a and b keeping character c constant.

Second order partial correlation is also tested by ά ô test

$$t = \frac{r_{ab.cd} \sqrt{(n-4)}}{\sqrt{1 - r^2_{ab.cd}}} \text{ with (n-4) degrees of freedom}$$

where, $r_{ab.cd}$ is the second order partial correlation.

In general, k^{th} order partial correlation is tested by:

$$t_{cal} = \frac{r_{12.34....(k+2)} \sqrt{n-k-2}}{\sqrt{1 - r^2_{12.34.....k+2}}} \text{ with (n-k-2) degrees of freedom}$$

If $|t_{cal}|$ is > t_{tab} at a specified Ŭ, then we reject the null hypothesis.

Multiple Correlation: Multiple correlation is an extension of the technique of simple correlation to the problems which involve two or more independent variables. Multiple correlation may be defined as a statistical tool designed to measure the degree of

relationship of one dependent variable with three or more independent variables. It is denoted by a symbol R. In a trivariate distribution, in which each of the variable X_1, X_2 and X_3 has n observations, the multiple correlation coefficient of X_1 with X_2 and X_3 is usually denoted by $R_{1.\ 23}$. The subscripts to the left of the dot stand for the dependent variable while the subscripts to the right of the dot represent the independent variables. The coefficient of multiple correlation can be expressed in terms of r_{12}, r_{13} and r_{23} as follows:

$$R_{1.23} = \sqrt{(r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23})/(1 - r^2_{23})}$$

$$R_{2.13} = \sqrt{(r^2_{12} + r^2_{23} - 2r_{12}r_{23}r_{13})/(1 - r^2_{13})}$$

$$R_{3.12} = \sqrt{(r^2_{31} + r^2_{32} - 2r_{31}r_{32}r_{12})/(1 - r^2_{12})}$$

where r_{12}, r_{13}, r_{23} stands for zero order or simple correlation coefficients. Multiple correlation coefficient can never be negative. It is necessarily positive or zero. By squaring $R_{1.23}$, we obtain the coefficient of determination which indeed is the per cent variability explained in dependent variable because of the influence of independent variables.

Testing the Significance of an observed Multiple Correlation Coefficient:

If R is the observed multiple correlation coefficient of a variate with ókô other variables in a random sample of size ónô from (k+1) variates population, then under the null hypothesis (H_o) that the multivariate correlation coefficient (R) in the population is zero, the statistic:

$$F = \frac{R^2}{1 - R^2} \quad \frac{n - k - 1}{k}$$ follows F distribution with (k, n-k-1) degrees of freedom.

If $F_{cal} \geq F_{Ŭ}(k, n ï\ k ï\ 1)$, then we reject H_0.

Example-7: In a poetry recitation competition, ten participants were ranked by two judges as:

Participant No.:	1	2	3	4	5	6	7	8	9	10
Judge x :	7	9	6	5	3	4	8	10	2	1
Judge y:	10	9	3	5	6	7	2	8	1	4

Measure the strength of relationship in the ranking behaviour of the two judges.

Solution: Let R_x and R_y denote the ranks given by the judges X and Y respectively.

Participant No.	$\mathbf{R_x}$	$\mathbf{R_y}$	$\mathbf{d = R_x - R_y}$	$\mathbf{d^2}$
1	7	10	-3	9
2	9	9	0	0
3	6	3	3	9
4	5	5	0	0
5	3	6	-3	9
6	4	7	-3	9
7	8	2	6	36
8	10	8	2	4
9	2	1	1	1
10	1	4	-3	9
				$\mathbf{\Sigma d^2 = 86}$

Here, n = 10 and $\Sigma d^2 = 86$

Using Spearmanôs rank correlation formula, we have

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 86}{10(10^2-1)} = 1 - \frac{6 \times 86}{10 \times 99} = 0.48$$

Testing significance of rank correlation coefficient: For n > 20 the distribution of Spearmanôs coefficient r_s tends to be normal with var $(r_s) = \frac{1}{n-1}$. However, for lower values of n (say 10) the statistic $t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$ follows (approximately) Studentôs t-distribution with (n-2) degrees of freedom.

Testing $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$

$$t_{cal} = \frac{r_s}{\sqrt{1-r_s^2}}\sqrt{n-2} = \frac{0.48}{\sqrt{1-0.48^2}}\sqrt{10-2} = 1.54$$

Table t-value at $\alpha = 0.05$ (Two-tailed test) for 8 df is 2.31

Since, $|t_{cal}| < t_{tab}$, hence the marks obtained by students of this group are independent.

Example-8: The marks of eight students in two papers economics and statistics are given below. Compute the rank correlation coefficient:

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Marks in Economics (X):	25	30	38	22	50	70	30	90
Marks in statistics (Y) :	50	40	60	40	30	20	40	70
Ranks assigned to X :	2	3.5	5	1	6	7	3.5	8
Ranks assigned to Y :	6	4	7	4	2	1	4	8
Difference in ranks (d) :	-4	-0.5	-2	-3	4	6	-0.5	0
d^2 :	16	0.25	4	9	16	36	0.25	0

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2-1)}$$

Here, $\Sigma d^2 = 81.5$, the item 30 is repeated twice hence $m_1=2$, item 40 is repeated thrice hence $m_2 = 3$.

$$r_s = 1 - \frac{6\left[81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right]}{8(8^2-1)}$$

$$= 1 - \frac{6[81.5 + 0.5 + 2]}{504} = 1 - 1 = 0$$

Hence, there is no rank correlation between the marks obtained in the two subjects.

Example-9: Calculate $r_{12.3}$, $r_{23.1}$ given r_{23}= -0.36, r_{31} = 0.34 and r_{12}= 0.70

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r^2_{13}}.\sqrt{1 - r^2_{23}}} = \frac{0.7 - (.34)(-.36)}{\sqrt{1 - (0.34)^2}.\sqrt{1 - (0.36)^2}} = 0.94$$

$$r_{23.1} = \frac{r_{23} - r_{12}.r_{13}}{\sqrt{1 - r^2_{12}}.\sqrt{1 - r^2_{13}}} = \frac{-0.36 - (0.7)(0.34)}{\sqrt{1 - (0.7)^2}.\sqrt{1 - (0.34)^2}} = -0.89$$

Example-10: Given the values n = 20, $r_{12.3}$ =0.7738, $r_{14.3}$ = 0.7243 and $r_{24.3}$ = 0.5262, calculate $r_{12.34}$ and test for its significance.

Solution: $r_{12.34} = \dfrac{r_{12.3} - r_{14.3}.r_{24.3}}{\sqrt{1-r^2_{14.3}}.\sqrt{1-r^2_{24.3}}}$

substituting the values, we get

$$r_{12.34} = \frac{.7738 - (.7243)(.5262)}{\sqrt{1-(.7243)^2}.\sqrt{1-(.5262)^2}} = 0.67$$

Test of significance: $H_0 : \rho_{12.34} = 0$ vs $H_1 : \rho_{12.34} \neq 0$ Let $\alpha = 0.05$

$$t_{cal} = \frac{r_{12.34}}{\sqrt{1-r^2_{12.34}}}\sqrt{n-4} = \frac{0.67\sqrt{20-4}}{\sqrt{1-0.67^2}} = \frac{0.67 \text{ x } 4}{0.74} = 3.62$$

Since, $|t_{cal}| = 3.62 > t_{0.05,16} = 2.12$, therefore, we reject H_0 and conclude that there is significant partial correlation between the variables.

Example-11: Calculate $R_{1.23}$ for the following data and test its significance

$n = 25$, $r_{12} = 0.60$, $r_{13} = 0.70$ and $r_{23} = 0.65$

Solution: We have $R_{1.23} = \sqrt{(r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23})/(1-r^2_{23})}$

$$= \sqrt{\{(0.6)^2 + (0.7)^2 - 2(0.6)(0.7)(0.65)\}/\{(1-(0.65)^2\}} = 0.725$$

Test of significance: H_o: There is no multiple correlation in the population

H_1: Multiple correlation in the population is greater than zero

Let Ŭ= 0.05

We know that:

$F = \dfrac{R^2}{1-R^2} \quad \dfrac{n-k-1}{k}$ follows F distribution with (k, n-k-1) degrees of freedom

So, $F_{cal} = \dfrac{(0.725)^2}{[1-(0.725)^2]} \text{ x } \dfrac{25-2-1}{2} = 12.19$

The tabulated value of F with (2, 22) d.f. at 5% level of significance is 3.44. Here the calculated value of F statistic is greater than the tabulated value. So we reject the H_0 and conclude that observed multiple correlation coefficient is significantly greater than zero.

EXERCISES

1. Obtain the coefficient of rank correlation between share and debenture prices.

Share Prices (X)	50	55	65	50	55	60	50	65	70	75
Debenture Prices (Y)	110	110	115	125	140	115	130	120	115	160

2. Calculate coefficient of simple correlation from the following data:

X:	11	8	4	2	6	7	8	6	8	4	10	8
Y:	10	3	1	2	3	5	2	10	11	3	2	5

3. Calculate the coefficient of rank correlation for the following data:

Students	1	2	3	4	5	6
Marks in math	75	40	52	65	60	80
Marks in statistics	30	45	35	50	48	42

4. The following table gives the frequency distribution according to age groups of marks obtained by 52 students in an intelligence test. Calculate the coefficient of correlation between age and intelligence.

Marks	**Age in years**				
	16-18	**18-20**	**20-22**	**22-24**	**Total**
10-20	2	1	1	-	4
20-30	3	2	3	2	10
30-40	3	4	5	6	18
40-50	2	2	3	4	11
50-60	-	1	2	2	5
60-70	-	1	2	1	4
Total	**10**	**11**	**16**	**15**	**52**