CORRELATION ANALYSIS

INTRODUCTION

In the previous Semester we have studied the characteristics of only one variable in the form of the measures of central tendency, the measures of dispersion, etc. A distribution of the values of one variable is called *univariate distribution*. In this chapter, we shall deal with problems and methods related with the determination of relationship between two or more variables. A distribution of paired values of two or more variables is called *bivariate distribution* or *multivariate distribution*.

In a bivariate distribution we may be interested to find if there exists any relationship between two variables such as income and expenditure, price and demand, height and weight, rainfall and crop yield etc. A statistical technique that is used to analyse the strength and direction of the relationship between two quantitative variables is called correlation analysis. It involves various methods and techniques used for studying and measuring the nature and degree of relationship between two or more variables. According to **A.M. Tuttle, "*Correlation is an analysis of the covariance between two or more variables*".**

MEANING OF CORRELATION

The term '*correlation*' means the quantitative relationship between two variables. Two variables are said to be correlated when the value of one variable changes with the change in the value of other variable. According to **Corxton and Cowden, "*When relationship is of a quantitative nature the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation*".**

The concept of correlation has been similarly defined by **W.I. King** as, "***If it is proved true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions, we consider that the fact is established and that a relationship exists. The relationship is called correlation*".**

Similarly, **Prof. Boddington** states that, "***Whenever some connection exists between two or more groups, classes or series of data, they are said to be correlated*".**

TYPES OF CORRELATION

The main types of correlation are as follows :

1. POSITIVE AND NEGATIVE CORRELATION

(*a*) ***Positive or Direct Correlation*** : If the values of two variables move in same direction, the correlation is said to be positive. In other words, if an increase or decrease in the values of one variable is associated with an increase or decrease in the values of the other variable, the correlation between them is said to be direct or positive. Let there be two variables X and Y. If

$\left.\begin{matrix} X\uparrow \rightarrow Y\uparrow \\ X\downarrow \rightarrow Y\downarrow \end{matrix}\right\}$ then X and Y are positively correlated. Some examples of positive correlation are :

(*i*) Price and supply of a commodity

(*ii*) Income and expenditure of a family

(*iii*) Height and weight

(*iv*) Temperature and sale of ice-cream during summer etc.

(*b*) ***Negative or Inverse Correlation*** : If the values of two variables move in the opposite direction, the correlation is said to be negative. In other words, if an increase or decrease in the value of a variable is associated with a decrease or increase in the values of the other, the correlation between them is inverse or negative. If

$\left.\begin{matrix} X\uparrow \rightarrow Y\downarrow \\ X\downarrow \rightarrow Y\uparrow \end{matrix}\right\}$ then X and Y are negatively correlated. Some examples of negative or inverse correlation are :

(*i*) Price and demand for a commodity

(*ii*) Number of workers and time required to complete the work

(*iii*) Volume and pressure of gas etc.

(*c*) ***Zero Correlation*** : Two variables are said to have zero correlation if they are not related with each other. In other words, if two variables are independent of each other then there is no or zero correlation between them. For example, the height of students and marks obtained by them, price of rice and demand for coffee have zero correlation.

2. SIMPLE PARTIAL AND MULTIPLE CORRELATION

(*a*) ***Simple Correlation*** : In simple correlation, the study relates to two variables only. For example, the study of correlation between two variables only, such as income and saving, price and demand etc.

(*b*) ***Partial Correlation*** : Under partial correlation, there are more than two variables and we study the relationship between any two variables keeping all other variables as constant. For example, studying the relation between yield of some crop (X) and chemical fertilizers (Y) without considering the effect of rainfall (Z) is known as partial correlation. Partial correlation between X and Y excluding Z will be represented by $r_{XY\cdot Z}$. Similarly $r_{YZ\cdot X}, r_{ZX\cdot Y}$ will be the symbols for other partial correlation coefficients.

(*c*) ***Multiple Correlation*** : If there are more than two variables and one variable is related to a number of variables, the study of relationship between one variable and all other variables taken together is called multiple correlation. For example, the study of relationship between production of a crop (X) and rainfall (Y), use of fertilizer (Z) taken together falls under multiple correlation. Multiple correlation coefficient is represented by $R_{x.yz}$.

3. LINEAR AND NON-LINEAR CORRELATION

(*a*) ***Linear Correlation*** : Two variables are said to be linearly correlated if the ratio of change between two variables is same or constant throughout the distribution. For example,

X :	1	2	3	4	5
Y :	3	5	7	9	11
Change in Y :		2	2	2	2

When these pair of values of X and Y are plotted on a graph, the line joining these points would be a straight line.

(*b*) ***Non-Linear Correlation*** : Two variables are said to have non-linear correlation if the ratio of change between two variables is not same or constant. Non-linear correlation is also called *curvilinear* correlation. For example :

X :	4	8	12	16	20
Y :	5	7	10	17	26

When these pair of values of X and Y are plotted on a graph paper, the line joining these points would not be a straight line.

4. LOGICAL AND ILLOGICAL CORRELATION

(*a*) ***Logical Correlation*** : It means when correlation between two variables is found out mathematically and this relationship is logical too. Correlation between demand and price, income and expenditure, yield of a crop and use of fertilizer are examples of logical correlation.

(*b*) ***Illogical Correlation*** : There are some cases when we come across with some variables which have no logical relationship with each other but mathematically we can establish a relationship between those by applying usual formulae of correlation analysis. For example, income and height of a group of persons. These variables are not related to each other in any way but correlation between them can be determined. Such a correlation is known as '*illogical correlation*' or '*non-sense correlation*' or '*spurious correlation*'.

DEGREES OF CORRELATION

The nature, extent or degree of relationship between two variables is studied with the help of correlation coefficient. On the basis of this degree, the nature and intensity or extent of correction is of following types :

(*i*) **Perfect Correlation** : Correlation is said to be perfectly positive if the value of coefficient of correlation is + 1 and perfectly negative if the value of coefficient of correlation is –1.

(*ii*) **No Correlation** : If there exists no relation between X and Y variables i.e. no interdependence exists between values of X and Y then correlation will not exist and it will be zero *i.e.* $r_{XY} = 0$.

(*iii*) **Limits for Degree of Correlation** : The limits of correlation as perfect, high, moderate or low in accordance with the values of the coefficient of correlation are given as under :

Degrees of Correlation	Positive Limits	Negative limits
(*i*) Perfect correlation	1	–1
(*ii*) Very high degree of correlation	0.9 or more	–0.9 or more
(*iii*) Fairly high degree of correlation	0.75 to 0.9	–0.75 to – 0.9
(*iv*) Moderate degree of correlation	0.50 to 0.75	– 0.50 to – 0.75
(*v*) Low degree of correlation	0.25 to 0.50	–0.25 to –0.50
(*vi*) Very low correlation	Less than 0.25	less than –0.25
(*vii*) Absence of correlation	0	0

USES OR SIGNIFICANCE OF CORRELATION

In real life, the study of correlation is significant in following ways :

(*i*) **It Reduces the Range of Uncertainty :** Prediction and forecasting plays an important role in policy making and planning. The study of correlation comes to our rescue in making relatively more reliable and dependable predictions.

(*ii*) **It Depicts the Average of Relationship :** While studying the movements in values of the variables we do not find, in general, any uniformity in it. Correlation analysis gives us a single value which can conclude the nature and extent of relationship between the variables.

(*iii*) **It Acts as Base for Other Statistical Measures :** Correlation analysis is closely related with regression analysis. With the help of regression analysis, we can estimate the value of one variable given the value of another variable. To sum up, we can quote **W.A. Neiswanger** that ***"Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the path through which stabilizing forces may become effective".***

(*iv*) **Helpful to Economists :** Correlation analysis is helpful for economists, because with its help, they can judge about the dependence and relationship of two variables.

Methods of Correlation

1. Karl Pearson's Co-efficient of Correlation

Karl Pearson's coefficient of correlation between two variables X and Y is a numerical measure of linear relationship between them. It is defined as the ratio of covariance between X and Y to the product of the standard deviations of X and Y.
It is also known as Product Moment method.
It is denoted by 'r'.

a) From Actual mean:

$$r = \frac{Cov.(X,Y)}{\sigma_x \sigma_y} \quad \text{---(1)}$$

We can elaborate this formula:

$$\therefore \quad Cov(X,Y) = \frac{\Sigma X-\bar{X}\,(Y-\bar{Y})}{n} = \frac{\Sigma xy}{n}$$

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

$$\sigma_Y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{n}} = \sqrt{\frac{\Sigma y^2}{n}}$$

Where $X-\bar{X} = x$, $Y-\bar{Y} = y$

Camlin | Page
Date / /

Substituting these values in (I), we have

a) When deviations are taken from Actual mean.

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

Direct method:

b) Without using Actual mean & Std. Deviation

$$r = \frac{N\Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$\Sigma X, \Sigma Y$ = Sum of both X & Y seperate. , ΣXY = Sum up product of X&Y

ΣX^2 & ΣY^2 = Sum up Square of both X & Y seperate.

c) When deviations are taken over Assumed mean

$$r = \frac{N\Sigma dx dy - (\Sigma dx)(\Sigma dy)}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

where $dx = X - A$, $dy = Y - A$

Unit - 4.

Camlin | Page
Date / /

Probable Error

Probable Error denoted by P.E.(r). is used to measure the statistical significance or reliability and dependab of coefficient of correlation.

Karl Pearson's probable Error is:

$$P.E. = 0.6745 \frac{1-r^2}{\sqrt{N}} \quad \frac{2}{3}$$

Where P.E. = Probable Error
r = Co-efficient of correlation
N = Number of pairs of observations.

Correlation Analysis

Unit-4

Camlin | Page
Date / /

Rank Correlation (Spearman's Method)

This method was developed by British Psychologist Prof. Charles Edward Spearman in 1904. Rank correlation is used for measuring the relationship between two qualitative variables such as honesty, beauty, taste etc. which cannot measure quantitatively. This method is used when ordinal or rank are available. It is known as Spearman's Coefficient of correlation or Rank correlation. Here, we take the differences in ranks, squaring them, and find the aggregate of square diff

Rank correlation is denoted by r_k

$$r_k = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

or

$$= 1 - \frac{6\Sigma D^2}{N^3-N}$$

where D is difference in ranks

n = No. of pairs in Observations

r_k = rank coefficient correlation

The value of r_k falls b/w ± 1

In rank coefficient of correlation, there are three different cases.

Case-I When ranks are not given.

eg. 40. Pg.No - 6.44

Case II When ranks are given

eg:- 2.

Case III When ranks are equal (repeated or tied).

$$\rho_k = 1 - \frac{6\left(\Sigma D^2 + 1/12(m^3 - m) + 1/12(m^3 - m) + \ldots\right)}{N^3 - N}$$

m stands No. of items which have common rank.

Concurrent Coefficient of Correlation.

Camlin Page
Date unit-4

3. **Concurrent Deviation Method:**

This method helps to know the direction of change of X and Y variables. This is denoted by r_c

$$r_c = \pm\sqrt{\pm\left(\frac{2C-n}{n}\right)}$$

Where

$n = N - 1$ (Total No. of Signs)

$C =$ No. of concurrent deviations or No. of +ive / positive signs

$N =$ Total No. of items