CHAPTER 6

Correlation Analysis

OBJECTIVES

After studying the material in this chapter, you should be able to :

- *Understand the concept of correlation between two variables.*
- *Identify different types of correlation.*
- *Understand the notion of correlation coefficient.*
- *Interpret the value of coefficient of correlation.*
- *Discuss various methods of computing the correlation coefficient.*
- *Compute correlation coefficient for bivariate frequency distribution.*
- *Appreciate properties of correlation coefficient.*
- *Know the merits and demerits of different methods of studying correlation.*
- *Calculate probable error and interpret its value.*
- *Understand the concept of coefficient of determination.*

6.1 INTRODUCTION

Thus far we have examined numerical methods used to describe various characteristics of a univariate data, *i.e.*, the data involving only one variable. The reader may recall that in univariate data only one variable is associated with each unit of observation. However, we may have data in which more than one variable can be associated with each unit of observation. For example, for providing information about the marks obtained by the students of a class in two subjects, say Statistics and Economics, we can associate two variables, one representing the marks in Statistics and the other marks in Economics to each unit of observation, namely, a student in the class. When we have two variables for which values are being observed for each unit of observation, we say that we have *bivariate data*. In general, the study of those data which involve more than two variables are termed as *multivariate data*.

Two variables are said to be *correlated* if the change in one variable is accompanied by a change in the other. For example, if X represents the price of a product and Y represents the demand for that product, than we would expect large values of X to correspond to

small values of Y and small values of X to correspond to large values of Y. Hence we can say that price and demand of a product are correlated. *Correlation analysis* is a statistical procedure by which we determine the degree of association or relationship between two or more variables. That is, in correlation analysis, the purpose is to measure the strength or closeness of the relationship between the variables. For example, we might find a high degree of relationship between the price of a product and consumer demand for that product. Correlation is said to be *linear* if the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable.

In this chapter we shall consider the problem of measuring the linear relationship involving two variables only. The study of such a problem is called the *simple linear correlation*.

6.2 CORRELATION : SOME DEFINITIONS

In the following we shall give some important definitions of correlation.

1. *If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in the other (s), then they are said to be correlated.*
— **L.R. Conno**

2. *Correlation analysis attempts to determine the degree of relationship between variables.*
— **Ya Lun Cho**

3. *Correlation is an analysis of the covariation between two or more variables.*
— **A.M. Tut**

4. *Correlation analysis deals with the association between two or more variables.*
— **Simpson and Ka**

5. *When the relationship is of quantitative nature, the appropriate statistical tool discovering and measuring the relationship and expressing it in a brief formu known as correlation.*
— **Croxton and Cow**

6. *Correlation means that between two series or group of data there exists some co connection.*
— **W.I.**

7. *When a group of items are recorded with respect to the values of two distinct vari and it is found that pairs of values tend to be associated, the two variables are s be correlated.*
— **Wessel and**

6.3 TYPES OF CORRELATION

The following are different types of correlation :

(i) *Positive and Negative Correlation.*

(ii) *Simple, Partial and Multiple Correlation.*

(iii) *Linear and Non-linear Correlation.*

(i) Positive and Negative Correlation

The correlation between two variables is said to be **positive or direct** if an increase (or a decrease) in one variable corresponds to an increase (or a decrease) in the other

For example, if X represents the amount of money spent annually on advertising by a firm and Y represents the total annual sales, then we might expect an increase (or a decrease) in the advertising budget to be accompanied by an increase (or decrease) in the annual sales. Thus we can say that the correlation between the advertising budget and the total sales is positive.

The correlation between two variables is said to be **negative or inverse** if an increase (or a decrease) in one variable corresponds to a decrease (or an increase) in the other. For example, if X represents the price of a product and Y represents the demand for that product, then we would expect large values of X to correspond to small values of Y and small values of X to correspond to large values of Y. Hence we can say that the correlation between the price and demand of a product is negative.

(ii) Simple, Partial and Multiple Correlation

The study of simple, partial and multiple correlation is based upon the number of variables involved.

Simple Correlation : It involves the study of only two variables. That is, in simple correlation, we measure the degree of association or relationship between two variables only. For example, when we study the correlation between the price and demand of a product, it is a problem of simple correlation.

Partial Correlation : It involves the study of three or more variables, but consider only two variables to be influencing each other, the effect of other influencing variables being kept constant. Thus in partial correlation we measure the degree of relationship between the variable Y and one of the variables $X_1, X_2, \ldots, X_n$ with the effect of all the other variables removed. For example, if we consider three variables, namely yield of wheat, amount of rainfall and amount of fertilizers and limit our correlation analysis to yield and rainfall, with the effect of fertilizers removed, it becomes a problem relating to partial correlation only.

Multiple Correlation : It involves the study of three or more variables simultaneously. Thus in multiple correlation we measure the degree of relationship between the variable Y and all the variables $X_1, X_2, \ldots, X_n$ taken together, For example, if we study the relationship between the yield of wheat per acre and both amount of rainfall and the amount of fertilizers used, it becomes a problem relating to multiple correlation.

(iii) Linear and Non-linear Correlation

Linear Correlation : The correlation between two variables is said to be *linear* if the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable. For example, consider the following data:

X :	10	20	30	40	50
Y :	40	80	120	160	200

We observe that the ratio of changes between the two variables is same and hence the correlation between X and Y is linear. It may be remarked that if the values of two variables, which are linearly correlated, are plotted on a graph paper all the plotted points would be on a straight line.

Definition The quantitative measure of strength in the linear relationship between two variables is called the **correlation coefficient.** It is denoted by r.

Thus the correlation coefficient r measures the extent to which the points cluster about a straight line. The correlation coefficient ranges from +1 to – 1. If two variables have no linear relationship, the correlation between them is zero. Consequently, the more correlation differs from zero, the stronger the linear relationship between the two variables.

The following table shows degrees of correlation according to various values of r:

TABLE 6.1

Degree of Correlation	*Positive*	*Negative*
Perfect correlation	+ 1	- 1
Very high degree of correlation	+ 0.9 to + 1	- 0.9 to - 1
Fairly high degree of correlation	+ 0.75 to + 0.9	- 0.75 to - 0.9
Moderate degree of correlation	+ 0.50 to + 0.75	- 0.50 to - 0.75
Low degree of correlation	+ 0.25 to + 0.50	- 0.25 to - 0.50
Very low degree of correlation	0 to + 0.25	- 0.25 to 0
No correlation	0	0

6.6 METHODS OF STUDYING CORRELATION

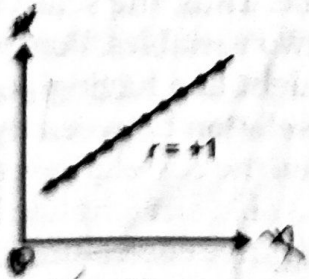
We shall discuss the following methods of measuring the linear relationship between two variables:

(i) *Scatter Diagram Method,*
(ii) *Karl Pearson's Coefficient of Correlation,*
(iii) *Rank Correlation Method, and*
(iv) *Concurrent Deviation Method.*

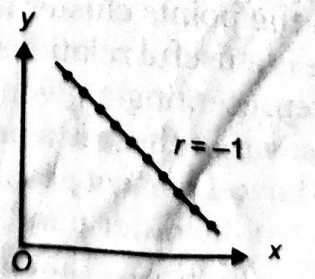
6.7 SCATTER DIAGRAM METHOD

A *scatter diagram* is a graphical presentation of bivariate data $\{(X_i, Y_i) : i = 1, 2, ..., n\}$ on two quantitative variables X and Y that allows us to show two variables together, one on each axis, each pair being represented by a point on the graph as in coordinate geometry.

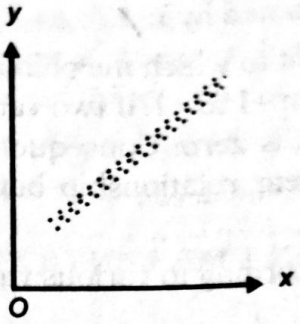
SCATTER DIAGRAMS SHOWING VARIOUS DEGREES OF CORRELATION



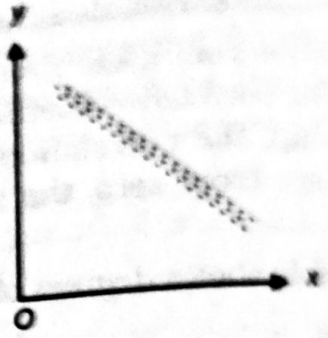
(a) Perfect positive correlation



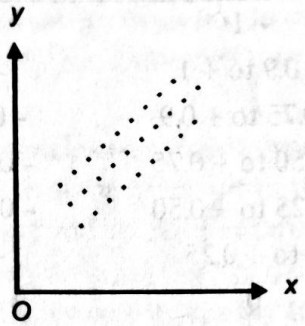
(b) Perfect negative correlation



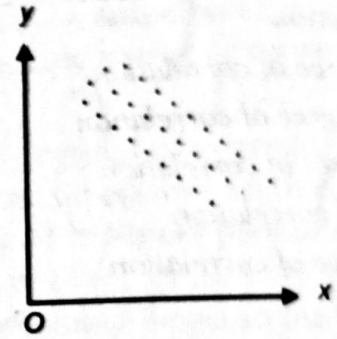
(c) *High degree of positive correlation*



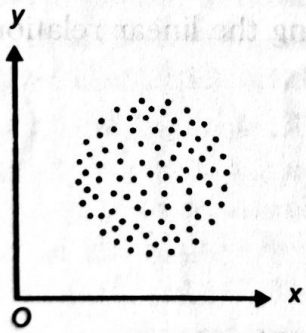
(d) *High degree of negative correlation*



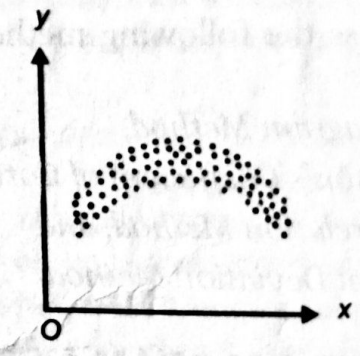
(e) *Low degree of positive correlation*



(f) *Low degree of negative correlation*



(g) *No correlation*



(h) *No correlation*

Fig. 6.1

The scatter diagram is the simplest method of measuring the linear relationship between two variables. By constructing a scatter diagram for the n pairs of observations (X_1, Y_1), (X_2, Y_2), . . . , (X_n, Y_n) on two variables, we can draw certain conclusions concerning the extent to which the points cluster about a straight line. Thus, the scatter diagram helps us to see if there is a useful relationship between the two variables. For example, if all the plotted points representing a given data lie on a straight line having positive slope [*see* Fig. 6.1 (*a*)], we say that there is a **perfect positive correlation** between the two variables. If two variables have a perfect positive correlation, then the correlation coefficient would be equal to + 1. On the other hand, if all the points lie on a straight line having negative slope [*see* Fig. (6.1(*b*)], we say that there is a **perfect negative correlation** between the two variables. If two variables have a perfect negative correlation, then the correlation

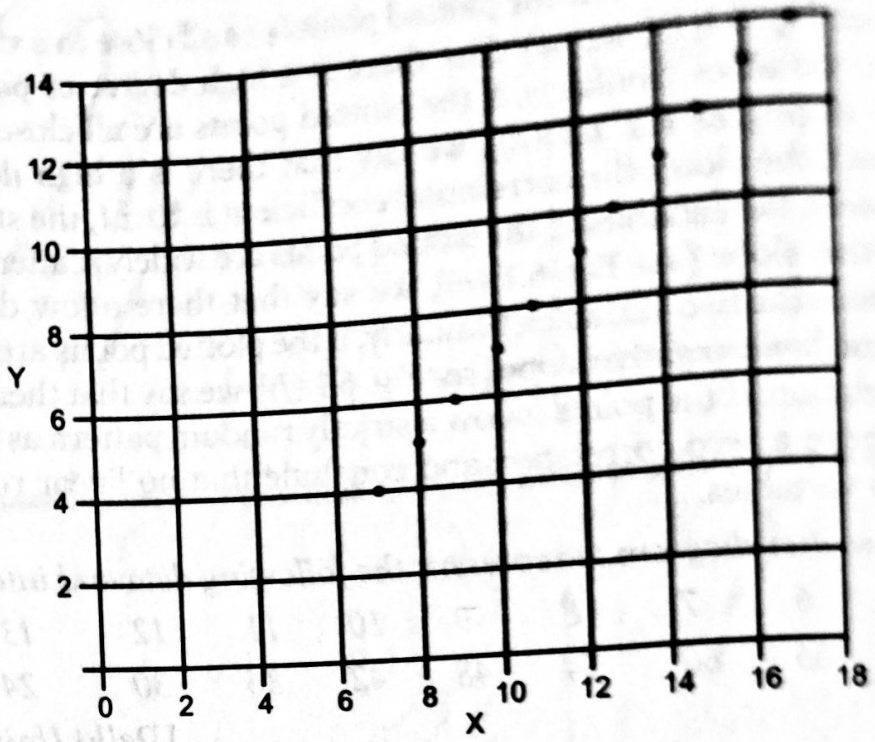


Fig. 5.3. Scatter Diagram

From the above scatter diagram, we find that the plotted points representing the given data lie on a straight line having positive slope. Hence we can conclude that there is a *perfect positive correlation* between the two variables.

6.8 COVARIANCE

In this section we introduce the concept of covariance between two quantitative variables. In the next section we shall see how this concept is used to measure the linear relationship between two variables.

Definition : Consider a set of n pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots\ldots (X_n, Y_n)$ on two quantitative variables X and Y, where $X_1, X_2, \ldots$ denote observed values of the variable X and $Y_1, Y_2, \ldots$ those of Y. The **covariance** between X and Y, denoted by Cov (X, Y), is defined as

$$\text{Cov}(X, Y) = \frac{(X_1 - \bar{X})(Y_1 - \bar{Y}) + (X_2 - \bar{X})(Y_2 - \bar{Y}) + \ldots + (X_n - \bar{X})(Y_n - \bar{Y})}{n}$$

$$= \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{n} = \frac{\Sigma xy}{n}$$

where $\bar{X} = \dfrac{\Sigma X}{n}$, $\bar{Y} = \dfrac{\Sigma X}{n}$, $x = X - \bar{X}$ and $y = Y - \bar{Y}$

EXAMPLE 8. *Find Cov (X, Y) between X and Y if*

X :	3	4	5	6	7
Y :	8	7	6	5	4

SOLUTION.

CALCULATION OF COVARIANCE

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	$xy = (X - \bar{X})(Y - \bar{Y})$
3	8	-2	2	-4
4	7	-1	1	-1
5	6	0	0	0
6	5	1	-1	-1
7	4	2	-2	-4
$\sum X = 25$	$\sum Y = 30$			$\sum xy = -10$

Here $n = 5$, $\bar{X} = \dfrac{\sum X}{n} = \dfrac{25}{5} = 5$ and $\bar{Y} = \dfrac{\sum Y}{n} = \dfrac{30}{5} = 6$

$\therefore$ $$\text{Cov}(X, Y) = \frac{\sum xy}{n} = \frac{-10}{5} = -2.$$

Another Formula for Cov (X, Y) : We now give a slightly different formula (proof omitted) for calculating the covariance. This formula is particularly useful when $\bar{X}$ or $\bar{Y}$ is not an integer. The formula is :

$$\text{Cov}(X, Y) = \frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right) \quad \ldots (1)$$

EXAMPLE 6. *Calculate the covariance between X and Y for the following data :*

X	:	1	2	3	4	5	6	7	8	9	10
Y	:	6	9	6	7	8	5	12	3	17	1

SOLUTION.

CALCULATION OF COVARIANCE

X	Y	XY
1	6	6
2	9	18
3	6	18
4	7	28
5	8	40
6	5	30
7	12	84
8	3	24
9	17	153
10	1	10
$\sum X = 55$	$\sum Y = 74$	$\sum XY = 411$

We have $n = 10$, $\sum X = 55$, $\sum Y = 74$ and $\sum XY = 411$

$$\text{Cov}(X, Y) = \frac{\sum uv}{n} - \left(\frac{\sum u}{n}\right)\left(\frac{\sum v}{n}\right) = \frac{12}{7} - \left(\frac{0}{7}\right)\left(\frac{-9}{7}\right) = 1.7.$$