UNIT 8 STATISTICAL TECHNIQUES

Structure

8.0 OBJECTIVES

After going through this Unit you should be in a position to:

- explain various measures of central tendency such as arithmetic mean, median and mode;
- explain various measures of dispersion such as range, variance, standard deviation, and coefficient of variation;
- explain correlation and regression techniques; and
- analyse time series data.

8.1 INTRODUCTION

In the previous Unit we explained the methods of presenting data in the form of tables and graphs. However, many times we need a single summary value that would describe a series. For example, we have data on the number of visitors to a library on a daily basis. Such data can be presented in the form of a table or in the form of a line graph. But if we want a single summary figure, arithmetic mean would give us the average number of visitors to the library.

Statistical techniques are more suitable for quantitative data although certain techniques do exist for qualitative data also. Recall that quantitative data can be of two types: discrete and continuous. We will explain various techniques for both types of variables. We begin with measures of central tendency.

8.2 MEASURES OF CENTRAL TENDENCY

Measures of central tendency provide us with a summary that describes some central or middle point of the data. There are five important measures of central tendency, viz., i) arithmetic mean, ii) median, iii) mode, iv) geometric mean, and v) harmonic mean. Out of these the last two measures, viz., geometric mean and harmonic mean, have very specific uses and thus less frequently used. Therefore, we will discuss the first three measures in this Unit.

Before dealing with these measures let us be familiar with certain notations that we will use. The standard notation is: X is a variable that takes values $X_1, X_2, X_3 \cdots X_N$. Let us consider the data given in Unit 7, Table 7.3 on the number of books issued to borrowers. You know that it is a discrete variable and the number of books that can be issued to each borrower varies between 0 and 5, viz., 0, 1, 2, 3, 4, and 5. The corresponding frequency for each observation is: 10, 23, 25, 17, 15 and 10.

In the above example, we denote the variable 'number of books issued' as X and the values assumed by it as $X_1, X_2, X_3, X_4, X_5, X_6$. The corresponding frequencies are $f_1, f_2, \cdots, f_6$. We call a typical observation as the i^{th} observation and denote it as X_i with frequency f_i. In the example on the number of books issued i ranges between 0 and 5.

In the case of continuous variable we take the mid-values of class intervals as $X_1, X_2, X_3 \cdots X_n$ and the corresponding frequencies as $f_1, f_2, \cdots, f_N$.

8.2.1 Arithmetic Mean

Arithmetic mean is also called 'mean' or 'average'. It is denoted by a bar above the variable being averaged. It is defined as the sum of all observations divided by the number of observations.

Let us calculate arithmetic mean for observations arranged in a frequency distribution. If $X_1, X_2, X_3 \cdots X_N$ are the observations and the corresponding frequencies are $f_1, f_2, \cdots, f_N$ then arithmetic mean is given by $\overline{X}$ (Read it as 'X-bar') and defined as

$$\overline{X} \quad \frac{1}{N} f_1X_1 \quad f_2X_2 \quad \cdots \quad f_nX_n \ .$$

It can be abbreviated as $\bar{X} = \frac{1}{N} \sum f_i X_i$ ...(8.1)

where N is the total number of observations and is equal to $\sum f_i$. The symbol $\sum$ (read it as 'sigma') denotes the sum of a variable.

When observations are classified into class intervals, in the case of continuous data, individual observations within a class interval are not separately identifiable. To avoid this difficulty, it is assumed that every observation falling into a class interval has a value equal to the *mid-value* of the class interval.

Example 8.1

Calculate the average number of books issued to borrowers on the basis of the following data.

Number of books issued	Number of borrowers
0	10
1	23
2	25
3	17
4	15
5	10
Total	100

We prepare a table for calculating arithmetic mean.

Number of books issued (X_i)	Number of borrowers (f_i)	$f_i X_i$
0	10	0
1	23	23
2	25	50
3	17	51
4	15	60
5	10	50
Total	$\sum f_i = 100$	$\sum f_i X_i = 234$

Secondly, we fill in the values from the table in the formula (8.1) that is $\bar{X} \quad \frac{1}{N} \quad \sum f_i X_i$.

Here $N = 100$, and $\sum f_i X_i = 234$.

Therefore, $\bar{X} \quad \frac{1}{100} \quad 234 \quad 2.34$

Thus, the average number of books issued to borrowers is 2.34.

Example 8.2

Given below is the monthly expenditure in Rupees on purchase of books by 100 persons. What is the monthly average expenditure?

Class Interval	Frequency
100-200	12
200-300	18
300-400	28
400-500	19
500-600	13
600-700	7
700-800	3
Total	100

In the case of continuous data we have to find out the mid-value of the class intervals and then apply formula (8.1).

Class Interval	Mid-value	Frequency	$f_i X_i$
100-200	150	12	1800
200-300	250	18	4500
300-400	350	28	9800
400-500	450	19	8550
500-600	550	13	7150
600-700	650	7	4550
700-800	750	3	2250
Total		$f_i = 100$	$f_i X_i = 38600$

By applying the relevant values from the above table in (8.1), that is,

$$\bar{X} \quad \frac{1}{N} \quad f_i X_i$$

Here $N = 100$, and $f_i X_i = 38600$.

Therefore, $\bar{X} \quad \frac{1}{100} \quad 38600 \quad 386$

Thus the average monthly expenditure on purchase of books by the groups of individuals is Rs. 386.

8.2.2 Median

Median gives us with the middle-most observation in a series so that half of the observations remain on each side of the median. For example, if you have 5 observations, viz., 2, 5, 9, 14 and 20, then 9 is the middle observation and 2 observations remain on both sides of it. Thus median of the above series is 9. Let us consider

another series where there are 6 observations: 3, 8, 15, 25, 35, and 43. In this case the median could be any number between 15 and 25 because 2 observations will remain on both sides. Conventionally we take the average of the middle-most two numbers. Here it would be $\frac{15 \quad 25}{2}$ 20. Thus in this case the median is 20.

However, when the number of observation is too large or data is arranged in a frequency distribution, it is not that simple to locate the median. If there are N observations the median observation should correspond to the th observation. We first find out the cumulative frequency of the distribution (see Unit 7) and secondly find out the class interval in which the $\frac{N}{2}$ th observation lies. This class interval is our 'median class'. Thirdly, we apply the following formula to get the median value.

$$M_d \quad l_m \quad \frac{\frac{N}{2} \quad C}{f_m} \quad h, \qquad \ldots(8.2)$$

where

l_m is the lower limit of the median class, i.e., the class in which median lies,

N is the total frequency,

C is the cumulative frequency of classes preceding the median class,

f_m is the frequency of median class, and

h is the width of median class.

$\frac{N}{2}$

Example 8.3

For the data given in Example 8.2 above, find out the median monthly expenditure on purchase of books.

To solve the above problem we go by the following steps:

1) calculate the cumulative frequency distribution

2) find out the median class

3) apply formula (8.2)

Class Interval	Frequency	Cumulative Frequency
100-200	12	12
200-300	18	30
300-400	28	58
400-500	19	77
500-600	13	90
600-700	7	97
700-800	3	100
Total	100	

There are 100 observation. So the median value corresponds to 50th observation, which lies in the class interval 300-400. Therefore, median will remain somewhere between 300-400. Thus the median class is 300-400. In this case

l_m =300, C = 30, N = 100, f_m = 28, h = 100,

By applying (8.2) we obtain the median value as

$$M_d \quad 300 \quad \frac{100/2 \quad 30}{28} \quad 100 = \text{Rs. } 371.43$$

8.2.3 Mode

Mode is the observation with the highest frequency. For discrete data it is easier to find out the mode. But in the case of continuous data we have to identify the 'modal class', that is the class interval having highest frequency. We have to see that the width of the classes is the same. Otherwise, large class intervals are likely to include large number of observations and smaller class intervals are likely to have few observations. Mode is computed by the following formula:

$$M_o \quad l_m \quad \frac{1}{1 \quad 2} \quad h, \qquad \ldots(8.3)$$

where

l_m is the lower limit of the modal class, i.e., the class in which mode lies,

$_1 \quad f_m \quad f_{m\ 1}$ is the difference of the frequencies of the modal class and its preceding class,

$_2 \quad f_m \quad f_{m\ 1}$ is the difference of the frequencies of the modal class and its following class, and

h is the width of modal class.

Example 8.4

Calculate the mode for the data set given in Example 8.2. The steps involved are

1) Identify the modal class. This is the class interval with highest frequency. In this case the modal class is 300-400.

2) Calculate l_m, $_1 \quad f_m \quad f_{m\ 1}$, $_2 \quad f_m \quad f_{m\ 1}$, and h

3) Apply formula (8.3)

We find that mode is $M_o \quad l_m \quad \frac{1}{1 \quad 2} \quad h \quad 300$.

$$= 352.63$$

Note that mean, median and mode assume different values for the same data. In the case of data relating to monthly expenditure on purchase of books given in Example 8.2, we find that mean, median and mode are Rs. 386, Rs. 371.43 and Rs. 352.63 respectively.

8.3 MEASURES OF DISPERSION

Measures of central tendency provide us with a summary figure for the data set. However, in many situations these measures do not represent the distribution of data. For example, look into the following three sets of data:

1) Set A: 2, 5, 17, 17, and 44.

2) Set B: 17, 17, 17, 17, and 17.

3) Set C: 13, 14, 17, 17, and 24.

Calculate the mean, median and mode for all three sets and you will find that they are the same, that is, 17 in all three sets. Still these sets are so different! While in Set B all the observations are equal, in Set A they are so dispersed. Definitely we need another measure, which will account for such dispersion of data.

The word dispersion gives the degree of heterogeneity in the data. It is an important characteristic indicating the extent to which observations vary amongst themselves. The dispersion of a given set of observations will be zero, only when all of them are equal as in Set B given above. The wider the discrepancy from one observation to another, the larger would be the dispersion. Thus dispersion in Set A should be larger than that in Set C. A measure of dispersion should capture such variability in data.

There are quite a few measures of dispersion. We will discuss range, mean deviation, variance and standard deviation in this Section.

8.3.1 Range

Range is defined as the difference between the largest and the smallest observations. Thus for the data given at Set A, the range is 44 - 2 = 42. Similarly, for Set B the range is 17 - 17 = 0 and for Set C it is 11. In the case of grouped data individual observations are not identifiable. In such cases we take the difference between two extreme boundaries of the classes.

8.3.2 Variance

Variance is the most widely used measure of dispersion. It is denoted by the symbol ^{2} (read as 'sigma-squared') and is defined as

$$\text{Variance} = {}^2 \quad \frac{1}{N} \quad (X_i \quad \bar{X})^2 \qquad \ldots(8.4)$$

In the case of frequency distribution variance is given by

$$^2 \quad \frac{1}{N} \quad f_i(X_i \quad \bar{X})^2 \qquad \ldots(8.5)$$

where $N \quad \sum_{i \ 1}^{n} f_i$, the total number of observations.

In order to simplify calculation we use the following formula

$$^2 \quad \frac{1}{N} \quad f_i X_i^2 \quad \bar{X}^2 \qquad \ldots(8.6)$$

Remember that we obtain the same value whether we apply (8.5) or (8.6).

8.3.3 Standard Deviation

Standard deviation is another widely used measure of dispersion. It is defined as the positive square root of variance and denoted by . Remember that standard deviation cannot be negative.

Example 8.5

Given below is the range of marks obtained by students in a class. Find out the standard deviation.

Marks	Number of students
15-25	8
25-35	12
35-45	20
45-55	10
55-65	6
65-75	4
Total	60

In order to calculate the standard deviation we go by the following steps:

1) calculate the mid-values of the classes
2) calculate arithmetic mean
3) Apply formula (8.5) or (8.6) to find out variance
4) Prepare a table with the required columns. We have prepared one for applying (8.5).
5) Find out variance
6) Find out the positive square root of variance

Marks	Number of students f_i	Mid-value X_i	X_i	$(X_i \quad \overline{X})$	$(X_i \quad \overline{X})^2$	$f_i(X_i \quad \overline{X})^2$
1	2	3	4	5	6	7
15-25	8	20	160	-21	441	3528
25-35	12	30	360	-11	121	1452
35-45	20	40	800	-1	1	20
45-55	10	50	500	9	81	810
55-65	6	60	360	19	361	2166
65-75	4	70	280	29	841	3364
Total	60		2460			11340

In the above table we first find out the arithmetic mean, which is 41. Next we find out variance as

$$^2 \quad \frac{1}{N} \quad f_i(X_i \quad \overline{X})^2 = \frac{1}{60} \quad 11340 \quad 189$$

Hence standard deviation is

$$\sqrt{189} = 13.75 \text{ marks}$$

Remember that standard deviation is always expressed in the unit of measurement. It is not a pure number. The difficulty with variance is that it is expressed in the square of the unit of measurement, which does not have any meaning.

8.3.4 Coefficient of Variation

Many times we have to compare the variability among different series of data. If the units are measured in different units we obtain different values for standard deviation. For example, let us compare the economic status of households in two villages. The summary figures of monthly calorie intake of households are given below for the two villages.

	Village A	**Village B**
Number of households	817	561
Mean calorie intake	2417	2235
S. D. of calorie intake	418	232

The problem is to identify the village that has more inequality as far as calorie intake is concerned. We find that village A has higher mean calorie intake but has larger standard deviation and larger number of households compared to village B. Thus village A may actually have more number of poorer households than in village B. In order to compare such situations we use the coefficient of variation (c.v.). It is defined as percentage standard deviation per unit of mean, i.e.,

$$\text{c.v.} = \frac{}{\overline{X}} \quad 100 \qquad \ldots(8.7)$$

Since and have same unit of measurement, c.v. is a pure number and it is not affected by the choice of unit of measurement.

For village A, c.v. $= \frac{418}{2417} \quad 100 = 17.29$ and for village B, c.v. $= \frac{232}{2235} \quad 100 = 10.38$.

Since the coefficient of variation in village A is greater than the coefficient of variation in village B, the inequalities are greater in village A compared to village B.

8.4 CORRELATION

So far we have dealt with a single characteristic of data. But, there may be cases when we would be interested in analysing more than one characteristic at a time. For example, you may like to study the relationship between the age and the number of books a person reads. Such data, having two characteristics under study are called bivariate data. One of the measures to find out the extent or degree of relationship between two variables is correlation coefficient.

An analysis of the co-variation of two or more variables is usually called correlation. If two characteristics vary in such a way that movement in one is accompanied by movement in the other, these characteristics are correlated. For example, there is relationship between price and supply, income and expenditure, etc. With the help of correlation analysis we can measure in one figure the degree of relationship existing between two variables.

8.4.1 Scatter Diagram

If we are interested in finding out the relationship between two variables, the simplest way to visualise it is to prepare a dot chart called scatter diagram. Using this method, the given data are plotted on a graph paper in the form of dots. For example, for each pair of X and Y values, we put a dot and thus obtain as many point as the number of observations. Now, by looking into the scatter of various dots, we can ascertain whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship. Here are some illustrations of some correlations between two variables.

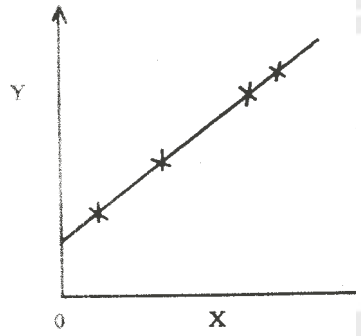

Fig. 8.1: Perfect Positive Correlation r = 1

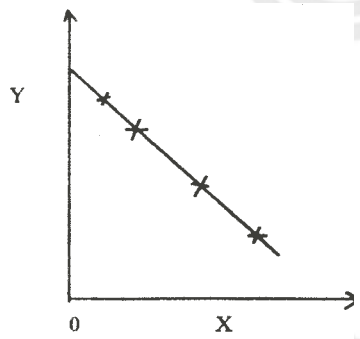

Fig. 8.2:Perfect Negative Correlation r = -1

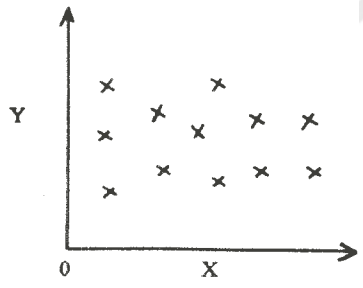

Fig. 8.3: No Correlation r = 0

You see that the scatter diagram gives a rough idea of the degree of relationship between two variables.

As we are considering a relationship between the two variables here, there might be a relationship between more than two variables. In case of 10 variables, we can plot the points on a two-dimensional graph paper, i.e., on a space with x and y axes. But the scatter diagram has the limitation that it cannot be plotted where more than two variables are involved. Secondly, it does not give an exact figure on the degree of relationship between variables.

To reach an exact figure on the extent of relationship between variables and also to overcome the limitation of considering more than two variables at a time, we calculate the correlation coefficient.

Fig. 8.4: High Degree of Positive Correlation

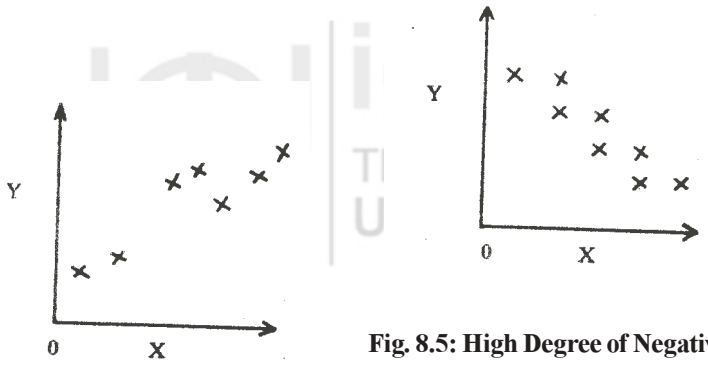

Fig. 8.5: High Degree of Negative Correlation

8.4.2 Pearson's Product Moment Correlation

Of the several methods of measuring correlation, Pearson's Product Moment correlation is mostly used in practice. It is denoted by the symbol r.

The formula is

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}}{\sqrt{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i\right.}}$$

Another formula for correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sigma_X \sigma_Y}$$

Where $_x$ is the standard deviation of X and $_y$ is the standard deviation of Y.

When we get value of r = +1, it means there is perfect positive correlation between variables. When r = -1, there is perfect negative correlation and when r = 0, it means that there is no correlation between the two variables. Correlation coefficient can take any value between +1 and -1, i.e., it cannot exceed +1 and cannot be less than -1. Usually, in real life analysis, we get values, which lie between +1 and -1 such as +0.6, -0.5 etc. The above formula of r can be transformed into the following form, which is easier to apply.

$$r = \frac{\sum_{i\ 1}^{n} y_i}{\sqrt{\sum_{i\ 1}^{n} y_i^2}}$$

Where $x_i \quad (X_i \quad \bar{X})$ and $y_i \quad (Y_i \quad \bar{Y})$

Steps necessary for calculating Coefficient of Correlation are:

i) Take the deviation of X from the mean X and denote this by

ii) Square this deviation and obtain the total, i.e.

iii) Take the deviation of Y from the mean Y and denote this by n

iv) Square this deviation and obtain

v) Multiply x and y obtain the total i.e.

vi) Substitute the values of $x_i y_i$ $x_i^{\ 2}$ and in the above formula.

Remember that correlation indicates the degree of association between variables X and Y. It merely shows whether the variation in one variable is accompanied by the variation in the other or not. Correlation coefficient does not indicate a cause and effect relationship between the variables. We cannot say X causes Y or vice versa. Second thing to remember is that a r = +0.6 does not imply greater relationship than r = 0.6 The positive or negative sign indicates the direction of relationship. If r = +0.6, then it is very likely than an increase in X is accompanied by an increase in Y. On the other hand, r = -0.6 indicates the inverse relationship between X and Y. Thirdly, correlation coefficient is neutral to change in scale and origin. This means that in illustration -8, for example, if we divide the X variable by any figure, say, 5 and variable Y by 2 (or any other figure), we get the same value for r-value. This is called a change in scale. For instance, when you convert the height in cm. To that in inches, the correlation coefficient does not change. Similarly, if you change the origin, i.e., deduct or add certain figure to all the observations, the correlation coefficient does not change. You may find this out yourself by deducting 10 from X and 5 from Y and then calculating the correlation coefficient. Fourthly, correlation coefficient indicates the linear relationship between the variables. For higher order relationships, correlation coefficient does not reflect the proper degree of relationship. For example, for two variables, X and Y, if X = Y2 for every value of Y then r may turn out to be zero. But this does not mean that X and Y are not related. So that we can say is that when two variables are independent r = 0, but from the mere knowledge of r = 0 we cannot infer that the two variables are independent.

Let us take an illustration and find out the value of r.

Illustration 11

The following table shows the data on height and weight of 10 children. Find out the product moment correlation coefficient.

Height (in cm)	Weight (in Kg)	Height (in cm)	Weight (in Kg)
110	26	140	38
110	21	135	30
125	22	130	30
130	24	140	40
145	36	135	43

Let the height (in cm.) be termed X and weight is Y

	Y_i	x_i $(X_i \quad \bar{X})$	y_i $(Y_i \quad \bar{Y})$	$x_i y_i$	x_i^2	y_i^2
110	26	-20	-5	100	400	25
110	21	-20	-10	210	400	100
125	22	-5	-9	45	25	81
130	24	0	-7	0	0	49
145	36	15	5	75	225	25
140	38	1	7	70	100	49
135	30	5	-1	-5	25	1
130	30	0	-1	0	0	1
140	40	10	9	90	100	81
135	43	5	12	60	25	144
1300	310				645	1300

X_i

8.4.3 Spearman's Rank Correlation

In the previous Unit, while mentioning the types of data, we had distinguished between ratio scale and ordinal scale of measurement. Pearson's product moment correlation coefficient can be applied only when the data are measured in the ratio scale. But when, instead of actual magnitude of the observations, we have only the ranks, the Pearson's 'r' becomes inapplicable. In such cases the Spearman's rank correlation (r_s) is used. The method of calculating r, is quite simple. The first step is to identify the rank or the i^{th} observation in X and Y in their respective series of n items. The second step is to find out the difference between the rank in X and the rank in Y. Let this difference be termed D1. The third step is to apply the following formula.

$$r_s \quad 1 \quad \frac{6 \quad \overset{n}{} \quad D_i^2}{\underset{i \ 1}{n(n^2 \quad 1)}}$$

The Spearman's rank correlation also ranges from +1 to –1. Thus, positive values indicate direct relationship between the variables, while negative values indicate inverse relationship. The value r = 0 indicates absence of association between the variables.

One note of caution is that, Spearman's rank correlation should not be used just because it is easier to compute than Pearson's product moment correlation coefficient..r, is less interpretive than r. We cannot strictly say that the change in X is associated with proportionate change in Y because equal differences in ranks do not imply equal differences in the characteristics.

Illustration 12

The following data give rank of 12 journals by two different methods used to compute rank correlation coefficient.

Sl. No.	Rank according to the No. of publication X	Rank according to the No. of citation Y	D_i	D_i^2
1	1	12	-11	121
2	2	9	-7	49
3	3	6	-3	9
4	4	10	-6	36
5	5	3	+2	4
6	6	5	+1	1
7	7	4	+3	9
8	8	7	+1	1
9	9	8	+1	1
10	10	2	+8	64
11	11	11	0	0
12	12	1	+11	121
				D^2 416

Here

Therefore, $r_{s=1}$ - 1.454 = 0.454

Self Check Exercise

1) Find out the standard deviation of the following data.

Marks in statistics	No. of students
0-20	5
20-40	10
40-60	15
60-80	8
80-100	2

2) The following table shows the marks obtained by 10 students in statistics and mathematics. Find out the correlation coefficient.

Sl. No. of Students	Marks in Statistics	Marks in Mathematics
1	52	48
2	60	40
3	45	38
4	38	28
5	35	42
6	62	65
7	68	53
8	28	25
9	50	28
10	62	33

3) Write the formulae for the following concepts:

a) Standard Deviation

b) Variance

c) Product Moment Correlation Coefficient

d) Rank Correlation

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

.....

8.5 REGRESSION ANALYSIS

In the previous Unit it has been noted that a correlation coefficient does not reflect cause and effect relationship. The regression analysis, to be discussed in this Unit, seeks to dwell on such a theme. It assumes that one variable is the cause and other(s) the effect. In general terms, we can say, variables are of two types: independent variables and dependent variables. Independent variable is the cause and dependent variable is the effect.

Regression analysis is a statistical tool, which helps understand the relationship between variables and predicts the unknown values of the dependent variable from known values of the independent variable.

Let us assume that the number of books in circulation, in a library is related to the number of users. For example, it can be postulated that as the number of users increases, the number of books in circulation also increases. Here, the number of users is the independent variable and the number of books in circulation is the dependent variable. Let us denote the dependent variable as Y and the independent variable as X. In regression analysis, we gather data over a period of time or across units at a point of time. Let us assume that 'n' pairs of observations in X and Y are collected. The next step is to find out the relationship X and Y.

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equations. The simplest of these equations is the linear equation. This means that the relationship between X and Y is in the form of a straight line. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Now the question arises, "how do we identify the equational form?" There is no hard and fast rule as such. The form of equation depends upon the reasoning and assumptions made by the researcher.

However, the researcher may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the equational from. If the points are more or less in a straight line then linear equation is assumed. If the points are not in a straight line and are in the form of a curve, a suitable non-linear equation, which resembles the scatter, is assumed.

The researcher has to make another assumption: viz. identification of independent and dependent variables. The again depends upon the logic of the researcher and purpose of analysis: whether Y depends upon X or X depends upon Y. Thus, there may be two regression lines from the same data (a) when Y is assumed to be dependent upon X, this is termed 'Y on X' line, and (b) when X is assumed to be dependent upon Y, this is termed 'X on Y' line.

Let us take an example of a linear equation with Y as the dependent variable and X as the independent variable.

$$Y=3+2X$$

By taking in different values of X, we can determine the values of Y, e.g., when X=1, Y=5; when X=2, Y=7 and so on. If we plot these pairs of points (1,5) (2,7), etc. on a graph paper we get a straight line.

Generalising the above relationship, it can be said that a linear equation of Y on X take3s the form Y=a+bX, where a and b are constants. Similarly, non-linear equations can be specified in many forms. A simple example is Y=a+bX+cX2

8.5.1 Linear Regression

Let us consider the following data. The number of visitors and the number of books issued during weekdays in a week are given.

No. of visitors to a library (X)	6	2	10	4	8
No. of books issued (Y)	8	4	10	7	8

If we plot the data on a graph paper, the scatter diagram looks something like Fig. 8.6.

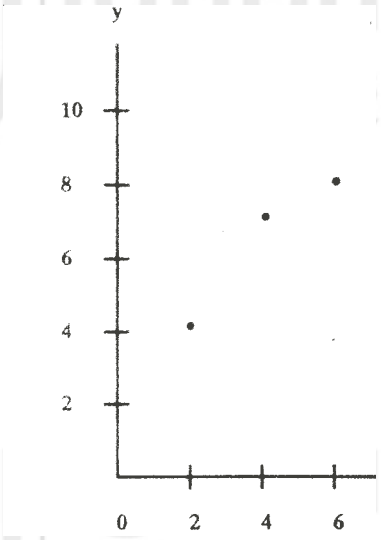

Fig. 8.6: Scatter Diagram for No. of visitors to a library and Books issued

As is obvious from the graph, the points do not strictly lie in a straight line. But they show an upward rising tendency where a straight line can be fitted.

If we plot the straight line along with the scattered points, the diagram looks like Figure 8.7. The difference between the regression line and the observations is the 'error'. For example, against a X value of 2, the Y value is 4. This is called the observed value.

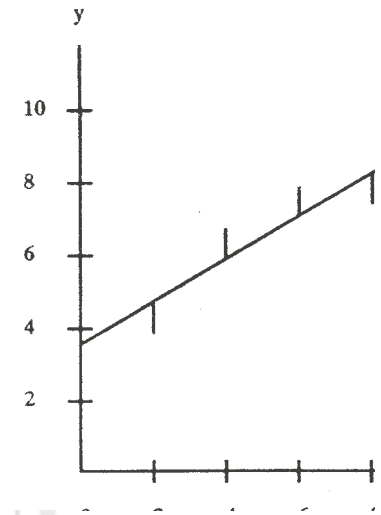

Fig. 8.7

But the regression line shows Y value of 4.8 against X value of 2. This value, which is calculated from the regression line, is the expected value. The difference between the observed value and the *expected value* is termed as the error value. So we see that observed value is the sum of expected value and error and value.

Our objective in fitting a regression line is to minimise the error values. This is usually done by the method of 'least squares'. The method of least squares minimises the value of E, where e is the difference between observed value and expected value. We will not go into the details of the method here. Instead, two equations derived on the basis of least squares method and known as normal equations are given.

These are:

$$Y \quad na \quad b \quad X$$

As a rule of thumb we can say that these normal equations are derived by multiplying the coefficients of 'a' and 'b' to the linear equation and summing over all observations. Here the linear equation is Y = a + bX. The first normal equation is simply the linear equation Y = a +bX summed over all observations.

SY = a +SbX or SY = na + bSX

The second normal equation is the linear equation multiplied by X and summed over all observations.

$SXY = SaX + SbX^2$ or $SXY = aSX + bSX^2$

It is evident that all the terms in these equations are given numbers, calculated from the data, except a and b.

The values of a and b need to be calculated for getting the estimated value of the dependent variable. This is done in the following.

It can be seen from Table 8.1 that data on X and Y variables are given in the first two columns. The succeeding two columns in the table give the calculations necessary to solve two normal equations given above. The expected value of Y and error value are given in the last two columns.

Table 8.1: Computations of Data for Regression Analysis

(1) X	(2) Y	(3) X2	(4) XY	(5) Expected value of Y	(6) Error
6	8	36	48	7.4	0.6
2	4	4	8	4.8	-0.8
10	10	100	100	10.0	0.0
4	7	16	28	6.1	0.9
8	8	64	64	8.7	-0.7
TOTAL 30	37	220	248	37	0.0

As the normal equations are:

Y = na + b X ……(8,8)

$XY = a \quad X + b \quad X^2$ ……(8.9)

We substitute the respective values from the Table 1.

Thus,

37 = 5a + 30b ……(8.10)

248 = 30a + 220b ……..(8.11)

If we multiply equation (3) by 6 and subtract the product from equation (4) we get :

248	=	30a	+	220b
222	=	30a	+	180b
-		-		-
		40b	=	26
or		b	=	0.65

On substituting the value of b in equation (8.10) we get:

37 = 5a + 30 x 0.65

or 5a = 17.5 or a = 3.5

So the regression line is

$$Y = 3.5 + 0.65 X \quad \ldots\ldots (8.12)$$

If we substitute the values of X in the regression line that is equation (8.12), we get the expected values of Y. For example, when X = 2

$$Y = 3.5 + 0.65 \times 2 = 4.8$$

But our observed value of Y against X = 2 is 4. This difference between the observed value and the expected value (4 - 4.8 = -0.8) is the error 'e'.

The expected values of Y and e are given in the Table 8.1 above in columns (5) and (6).

Notice that the sum of errors for the sample is zero, i.e. Se = 0

For computational purposes we may use the following formulae to find out the value of a and b.

Let us take

$$X = (X - X)$$

$$Y = (Y - Y)$$

$$XY = (X - X)(Y - Y)$$

Where $\bar{X}$ and $\bar{Y}$ are the arithmetic means of X and Y variables respectively. This formula gives

$$b \quad \frac{xy}{x^2} \quad \ldots\ldots (8.13)$$

$$a \quad \bar{Y} \quad b\bar{X} \quad \ldots\ldots (8.14)$$

Since these formulae are derived from the normal equations, we get the same values for 'a' and 'b' in this method also.

The steps in computation are:

1) Find out the values of X and Y

2) Find out x = (X - X) and y = (Y - Y)

3) Find out the values of xy

4) Find out the values of x^2

5) Apply the formulae in equations (8.13) and (8.14) above.

On applying the above formulae in the example given in Table 1, above we get the data given in Table 2.

Table 8.2: Computation of Regression Equation: Short cut Method

X	Y	X= (X - X)	Y = (Y - Y)	xy	X^2
6	8	0	0.6	0	0
2	4	-4	-3.4	13.6	16
10	10	4	2.6	10.4	16
4	7	-2	-0.4	0.8	4
8	8	2	0.6	1.2	4
Total 30	37			26.0	40

$\bar{X} \quad \frac{1}{n} \quad X \quad 30/5 \quad 6 \qquad b \quad \frac{xy}{x^2} \quad 26/40 \quad 0.65$

$\bar{Y} \quad \frac{1}{n} \quad Y \quad 37/5 \quad 7.4 \qquad a \quad \bar{Y} \quad b\bar{X} \quad 3.5$

Needless to mention that values for a and b are same as derived earler.

8.5.2 Properties of Regression Coefficient

Coefficient 'b' is called the regression coefficient. Notice that we can draw two regression lines from the data on X and Y.

(a) Y on X line, $Y = a + bX$

(b) X on Y line, $X = \quad + \quad Y$

The two coefficients, b and ß, demonstrate some interesting properties. First, the product of both regression coefficients is equal to the square of r (correlation coefficient), i.e., $b\ ß = r^2$

So once we know both regression coefficients we can find out the value of r^2. By taking the square root of r^2 we get r. Second if the regression coefficients are negative in sign, then the correlation coefficients also is negative. If the regression coefficients are positive then correlation is positive. Third, you know that

, i.e., r lies between - 1 and +1

Therefore r^2 lies between zero and +1. Regression coefficient can take finite value. But is one regression coefficient is more than 1 the other regression coefficient is less than 1. Both regression coefficients cannot exceed unity. Also it follows that the product of both, which is r^2, cannot exceed unity. The square of correlation coefficient is called the coefficient of determination and implies important characteristics. If r^2, the coefficient of determination, is closer to one we can infer that the independent variable explains the movements in the dependent variable. If the coefficient of determination is closer to zero, the independent variable does not explain the variation in the dependent variable.

8.5.3 Non-linear Regression

In the previous sub-section we discussed the simple linear regression involving two variables: one dependent and the other independent. Regression can involve one dependent variable and more than one independent variable. Such cases are called multiple regressions.

The equation fitted in regression can be non-linear or curvilinear. It can take numerous forms. A simpler form involving two variables is the quadratic form. The equation is

$Y = a + bX + CX^2$

There are three parameters here, viz., a,b and c, and the normal equations are :

$$Y = na + b\ X + c\ X^2$$

$$XY = a\ X + b\ X^2 + c\ X^3$$

$$X^2 Y = a\ X^2 + b\ X^3 + c\ X^4$$

Notice again that the normal equations are the regression equation multiplied by the coefficients of a,b and c and summed over all observations.

Certain non-linear equations can be transformed into linear equations by taking logarithms. Finding out the optimum values of the parameters from the transformed linear equations is the same as the process discussed in the previous sections. We give below some of the frequently used non-linear equations and the respective transformed linear equations.

1) $Y = a\ e^{bx}$

$\sqrt{\ }$ $\frac{1}{a\ \ bX}$ By taking natural log (that is, ln), it can be written as

$\ln Y = \ln a + b X$

Or $Y = a + ß X$

Where, $Y = \ln Y$, $a = \text{in } a$, $X = X$ and $ß = b$

2) $Y = a X^b$

By taking log, the equation can be transformed into

$\log Y = \log a + b \log X$

Or $Y = a + ß X$

Where, $Y = \log Y$, $= \log a$, $ß = b$ and $X = \log X$

3)

If we take $Y' \quad \frac{1}{Y}$ then

$Y' \quad a \quad bX$

4) $Y = a + b \quad X$

If we take $X = \quad X$ then

$Y = a + b X$

Once the non-linear equation is transformed, the fitting of a regression line is as per the method discussed in the beginning of this section. We derive the normal equations and substitute the values calculated from the observed data. From the transformed parameters, the actual parameters can be obtained by making the reverse transformation

8.5.4 Prediction

A major interest in studying regression lies in its ability to forecast. In the illustration at the beginning we assumed that the number of books issued depend upon the number of visitors. We fitted a linear equation to the observed data and got the relationship

$$Y = 3.5 + 0.65 X$$

From this equation we can forecast the number of books issued given the number of visitors. For example, if the number of visitors goes up to 30, then the number of books issued will be

$$Y = 3.5 + 0.65 \times 30 = 23$$

The procedure is to substitute the X value in the regression equation and get the expected Y value.

The question that arises here is: Will the predicted value come true? It depends upon the coefficient of determination. If the coefficient of determination is closer to one, there is greater likelihood that the prediction will be realized. However, the predicted value is constrained by elements of randomness involved with human behaviour and other unforeseen factors.

Self Check Exercise

4) Given below is the data on X and Y

X :	15	17	20	22	25	33
Y :	25	22	30	31	37	35

i) Find out the regression line Y on X.

ii) Find out the regression line X on Y.

iii) Find out the coefficient of determination.

iv) Find out the Pearson's product moment correlation.

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of the Unit.

.....

.....

.....

.....

.....

.....

.....

8.6 TIME SERIES ANALYSIS

In regression analysis, we discussed the cause and effect relationship between the dependent and the independent variables. The independent variable can be any characteristic based on our reasoning. When the independent variable is 'time', we call it 'time series'. We assume that the dependent variable depends upon time or varies according to time.

In our day-to-day life we encounter several instances where certain characteristics vary according to time. The stock in a library, the expenditure on user facilities, etc., are a few examples.

One of the important tasks before librarians and information managers is to make estimates for the future. For example, a publisher may want to know his probable sales for the next year, so that he can properly plan and take steps to avoid the possibility of unsold stocks or lack of supply of some books published by him. A librarian may wish to study the trend of book issued in order to take appropriate measures while making his future plans. For all these purposes, one needs to consult the data, which have been collected and recorded at successive intervals of time. Such statistical data is referred to as 'time series' data.

An example of time series data could be the number of books on Library and Information Science issued during 1995 to 2004. These data may be recorded as follows:

Table 8.3: Hypothetical Data Recorded on Books Issued Over Time

Year	No.of Books Issued
1995	356
1996	350
1997	391
1998	289
1999	408
2000	412
2001	405
2002	482
2003	497
2004	469

A close look at the data would show that the demand of the books has increased with some fluctuations. There may be several reasons for increase or decrease during a certain period.

8.6.1 Components of Time Series

The fluctuations in time series may be classified into the following types of variations.

a) Secular trend

b) Seasonal variation

c) Cyclical variation

d) Irregular variation

A time series may have the above components in combination as well.

a) **Secular Trend**

Changes which take place as a result of general tendency of the data to increase or decrease are known as secular movement. The general movement persisting over a long period of time is called secular trend. The above time series example for data on books issued is an example of secular trend.

b) **Seasonal Variation**

Changes or variations, which seasonally occur within a period of one year as a result of changes in climate, weather, important happenings etc., are called seasonal variations. It may be possible that in the data on books issued in the above example (If made available on a month to month basis), the year starts with a low figure in the beginning, reaches its peak in the middle and decreases at the end of the year. This type of Fluctuation within a span of a year is called seasonal variation.

c) **Cyclical Variation**

Changes that take place due to cyclic fluctuations like prosperity and depression may be termed as cyclical variations. In every business cycle there are four periods, (i) prosperity (ii) decline, (iii) depression and (iv) improvement. Cyclical variations are of a longer duration than a year.

d) **Irregular Variation**

Changes, which take place due to the factors that, could not be predicted like violent riots, natural calamities, etc., come under irregular variation.

The components of time series data, viz., seasonal, trend, cyclical and irregular, can be separated. In a traditional time series analysis, it is assumed that there is a multiplicative relationship among these four components. This may be represented symbolically as follows:

$$Y = T \times S \times C \times I$$

Where T = Secular Trend

S = Seasonal Variation

C = Cyclical Variation

I = Irregular Variation

This is called the multiplicative model. In another approach, it is assumed that

$$Y = T + S + C + I$$

A model formulation of this category is called 'additive model'.

8.6.2 Measurement of Secular Trends

There are various methods for determining secular trends.

The most frequently used methods are:

1. Moving average method, and
2. Method of least squares.

a) **Method of Moving Average**

It is a method of smoothing out fluctuations by calculating a series of averages by allowing overlapping periods of the time series. Before the moving average is calculated, it is necessary to select a proper period of moving average like three yearly, five-yearly, etc.

If the period chosen is m years, the moving averages are obtained by calculating a series of mean values of m consecutive values covering overlapping periods of the series. The mean of the first m values, given by $(Y_1 + Y_2 + \ldots\ldots + Y_m)$ is placed at the mid-point of the period covering the first m years.

It would be the first moving average value. The second moving average value will be obtained by calculating mean of values covering the period 2nd to (m+1)th years. For example: $\frac{1}{m}(Y_1 + Y_2 + \ldots\ldots + Y_{m\ 1})$ will be the next moving average and so on. This process is repeated till the last observation is covered.

The formula of the three-yearly moving average will be:

1/3 $(Y_1 + Y2 + Y_3)$, 1/3 $(Y_2 + Y_3 + Y_4)$, 1/3 $(Y_3 + Y_4 + Y_5)$ and so on.

And that of five-yearly moving average will be

1/5 $(Y_1 + Y_2 + Y_3 + Y_5)$, 1/5 $(Y2 + Y_3 + Y_4 + Y_5 + Y_6)$ and so on.

$\frac{1}{m}$

The above procedure will be easy to understand from the following illustration.

Illustration

Calculate a three-yearly moving average from the following sales figures of a publisher.

Table 8.4: Hypothetical Data for Computation of Moving Average

Year	Sales (hundred units)
1990	5
1991	7
1992	9
1993	12
1994	11
1995	10
1996	8
1997	12
1998	13
1999	17
2000	19
2001	14
2002	13
2003	12
2004	15

Data on sales of books for 15 years are given above. To calculate three-yearly moving average, we take, to begin with, the first three years total. This is 5+7+9=21 and place this value against the middle observation, i.e., against 1971. In column (4) of Table 8.5 the moving average, i.e., the three-yearly total divided by the period (which is 3 in this case) is given. Thus 21/3=7 for the first entry.

Following a similar procedure other calculations are made and results are presented in the table given below.

Table 8.5: Computation of Three-Yearly Moving Average

Year	Sales	3-yearly Totals	3-yearly Moving Average
(1)	(2)	(3)	(4)
1990	5		
1991	7	21	7.0
1992	9	28	9.33
1993	12	32	10.67
1994	11	33	11.00
1995	10	29	9.67
1996	8	30	10.00
1997	12	33	11.00
1998	13	42	14.00
1999	17	49	16.33
2000	19	50	16.67
2001	14	46	15.33
2002	13	39	13.00
2003	12	40	13.33
2004	15		

From the illustration above it may be noted that we do not get moving average value for the beginning and the end years. In the case of a five-yearly moving average, we lose moving averages for the beginning two years and the two years at the end. This loss of information increases as the 'period' of moving average increases. Secondly, from the moving averages, we cannot predict the figures for the future. It is just an analysis of past behaviour.

b) **Method of Least Squares**

We used this method earlier to obtain the regression lines. The procedure here is very similar to the fitting of regression lines. Here the independent variable is 'time' t. The first step is to form the equation of 'secular trend'. As you know, both straight lines and curves can be fitted by the least squares method. If Y is the dependent variable, the straight line to be fitted is $Y = a + bt$. Again, we have to minimize the error between the observed and the expected values. The method of least squares suggests that the sum

of squares of the error terms should be the minimum. From this method, the relationship between dependent and independent variables is estimated from the normal equations. For a linear equation Y = a +bt, the normal equations are:

$Y = na + b \quad t$

$tY = a \quad t + b \quad t^2$

The constants 'a' and 'b' are determined from these two equations, and 'n' indicates the number of observations in the sample.

We measure the variable 't' by taking the mid-point of time as the origin. Suppose n =5 years. Then taking the origin at the third year of the time, we get t = -2, -1, 0, 1, 2. It may be seen that t = 0 for the third year.

In case the number of years covered is even, say 6 years, the origin at the mid point of the two middle years, i.e., at 6 months past the third year is considered. In such cases t takes values

t = -5, -3, -1, +1, +3, +5

To fit a linear trend line by making use of the data given in Table 8.5 above, the necessary computations are summarised in Table 8.6 below.

Table 8. 6: Data Computation of Time Trend

Year	t	Sales	tY	t2
1990	-7	5	-35	49
1991	-6	7	-42	36
1992	-5	9	-45	25
1993	-4	12	-48	16
1994	-3	11	-33	9
1995	-2	10	-20	4
1996	-1	8	-8	1
1997	0	12	0	0
1998	1	13	13	1
1999	2	17	34	4
2000	3	19	57	9
2001	4	14	56	16
2002	5	13	65	25
2003	6	12	72	36
2004	7	15	105	
Total		0	177	171 280

Recall that the normal equations are:

$Y = na + b \quad t$

$tY = a \quad t + \quad t^2$

Substituting the respective values from the values from the table, we get,

177 = 15 a + b 0

171 = a 0 + b 280

or

15a =177 or, a = 11.8

280b =171 or, b = 0.61

So the trend line is Y = 1.8 +0.61 t

Remember that 't' is the codified time value with 197 as the origin.

The method of least squares enables us to forecast future values for Y. This is done by substituting the 't' value in the equation.

In the illustration given above, the sales books (in hundred units) in the year 2006 will be 17.29.

Y = 11.8 +0.61 9 = 17.29

In place of 't' 9 is substituted since starting with 1997, the year of origin, 2006will be 9 years. The predicted sales of books (in hundred units) in the year 2006 will be 17.29.

As you know, non-linear trends can also be fitted to the observed data. Hence, predictions can also be made on them. In the analysis of trend line, we make the implicit assumption that the past behaviour continues to persist in a future period also. So, a change in the past behaviour would make prediction unreliable.

8.6.3 Measurement of Other Components

Seasonal variation and cyclical fluctuation are periodic and recurring movements in the data. It has been stated above that the seasonal variation is short term in nature and usually the periodicity is less than a year. In contrast to this, the cyclical fluctuation lasts longer than a year.

Just as there are several methods of measuring the seasonal variations viz., ratio to trend method, ratio to moving averages method and link relative method, the cyclical fluctuations can be measured by harmonic analysis, spectrum analysis, etc. These methods involve tedious calculations and hence are not discussed here. If you are interested in these methods, you may look into the books referred to at the end of this Unit. However, attempts to separate the time series of its four components - seasonal, trend, cyclical and erratic, may follow some simple procedure. This is briefly spelt out in the following:

a) The seasonal component described above can be estimated with the help of moving average. This component can be eliminated from the original observations through subtraction if we assume an additive model..

b) The trend of the seasonally adjusted data are then estimated by means of least square straight line or some other function fitted by least squares described above. The trend component can be eliminated from the seasonally adjusted data.

c) The residuals, which remain after the elimination of seasonal and trend components from the original time series can be recorded and potted graphically. This residual variation may be compared visually or through some other method. The remaining variations of the data series are attributed to cyclical and erratic components.

8.7 SUMMARY

In this Unit we discussed various statistical techniques for analyzing data. Often it is necessary to provide a summary figure for a set or series of data. Such figure could be a measure of central tendency such as mean, median and mode, or it could be a measure of dispersion such as variance, standard deviation and coefficient of variation.

There are cases where more than one characteristic of a sampling unit is measured. In such type of data, we can find out the correlation coefficient or we can fit a regression equation. Remember that correlation does not show a cause and effect relationship between variables. It only shows the strength of relationship. In regression analysis variables are divided into two categories: independent and dependent. Regression equation can be a straight line or a curve depending upon the type of equation fitted.

Often we have data at certain intervals for a sufficiently long period of time. Such data re called time series and contains certain components, viz., secular trend, cyclical variation, seasonal variation and irregular movements.

8.8 ANSWERS TO SELF CHECK EXERCISES

1) The standard deviation comes out to be 21.07. Apply the formula given in the text and check the answer.

2) The correlation coefficient r = +0.61.

3) a) It is defined as the positive square root of variance and denoted by .

$$= +$$

^{2}

b) Variance is the most widely used measure of dispersion. It is denoted by the symbol ^{2} (read as 'sigma-squared') and is defined as

Variance = $^2 \quad \frac{1}{N} \quad (X_i \quad \bar{X})^2 \quad \frac{1}{N} \quad X_i \quad \bar{X}^2$

In the case of frequency distribution variance is given by

$$^2 \quad \frac{1}{N} \quad f_i(X_i \quad \bar{X})^2$$

where $N \quad \sum_{i \ 1}^{n} f_i$, the total number of observations.

In order to simplify calculation we use the following formula

$$^2 \quad \frac{1}{N} \quad f_i X_i^2 \quad \bar{X}^2$$

c) $$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sigma_X \sigma_Y}$$

d) $r_s \quad 1 \quad \frac{6 \sum_{i \ 1}^{n} D_i^2}{n(n^2 \quad 1)}$

4) i) You have to find out the normal equations and substitute the values in the equations. The estimated regression line will be Y = 13.94 + 0.73 X

ii) The regression line for X on Y will be X = a + bY

Consequently the normal equations will be

$\Sigma X = n\alpha + \beta \Sigma Y$

$\Sigma XY = \alpha \Sigma Y + \beta \Sigma Y^2$

The estimated regression line will be

X = -5.6 + 0.92 Y

iii) The coefficient of determination is the product of regression coefficients of both the regression lines. So, it is 0.73 x 0.92 = 0.067

iv) Correlation coefficient is the square root of coefficient of determination, i.e., r = Ö 0.067 = 0.82. Since the regression coefficient is positive in sign, the correlation coefficient is also positive.

8.9 KEYWORDS

Arithmetic Mean : Sum of observed values of a set divided by the number of observations in the set is called a mean or an average.

Median : In a set of observations, it is the value of the middlemost item when they are arranged in order of magnitude.

Mode : In a set of observations, it is the value which occurs with maximum frequency.

Coefficient of Variation : It is a relative measure of dispersion which is independent of the units of measurement. As opposed to this Standard Deviation is a pure number.

Range : It is the difference between the largest and the smallest observations of a given set of data.

Standard Deviation : It is the positive square root of the variance.

Variance : It is the arithmetic mean of squares of deviations of observations from their arithmetic mean.

Normal Equations : A set of simultaneous equations derived in the application of the least squares method, for example in regression analysis. They are used to estimate the parameters of the model.

Regression : It is a statistical measure of the average relationship between two or more variables in terms of the original units of the data.

Cyclical Variations : Oscillatory movements of a time series where the period of oscillation, called cycle, is more than a year.

Irregular Movement : The random movement of time series, which is not explained by other components. In this sense it is a residual of other components.

Method of Least Squares : When a polynomial function is fitted to the time series, the method of least squares requires that the parameters of the function should be so chosen as to make the sum of squares of the deviations between actual observations and expected values to be minimum.

Seasonal Variation : Periodical movement where the period is not longer than one year.

Secular Trend : The smooth, regular and long-term movement of a time series over a period of time. Trend may be upward or rising, downward or declining or it may remain more or less constant over time.

8.10 REFERENCES AND FURTHER READING

Sanders, D. H. (1980). *Statistics: A Fresh Approach.* New Delhi: McGraw Hill.

Rao, I. K. R. (1983). *Quantitative Methods for Library and Information Science.* New Delhi: Wiley Eastern.