12101CH02

2 Data Processing

You have learnt in previous chapter that organising and presenting data makes them comprehensible. It facilitates data processing. A number of statistical techniques are used to analyse the data. In this chapter, you will learn the following statistical techniques:

1. Measures of Central Tendency
2. Measures of Dispersion
3. Measures of Relationship

While measures of central tendency provide the value that is an ideal representative of a set of observations, the measures of dispersion take into account the internal variations of the data, often around a measure of central tendency. The measures of relationship, on the other hand, provide the degree of association between any two or more related phenomena, like rainfall and incidence of flood or fertiliser consumption and yield of crops.

Measures of Central Tendency

The measurable characteristics such as rainfall, elevation, density of population, levels of educational attainment or age groups vary. If we want to understand them, how would we do ? We may, perhaps, require a single value or number that best represents all the observations. This single value usually lies near the centre of a distribution rather than at either extreme. The statistical techniques used to find out the centre of distributions are referred as **measures of central tendency**. The number denoting the central tendency is the representative figure for the entire data set because it is the point about which items have a tendency to cluster.

Measures of central tendency are also known as statistical averages. There are a number of the measures of central tendency, such as the **mean**, **median** and the **mode**.

Mean

The mean is the value which is derived by summing all the values and dividing it by the number of observations.

Median

The median is the value of the rank, which divides the arranged series into two equal numbers. It is independent of the actual value. Arranging the data in ascending or descending order and then finding the value of the middle ranking number is the most significant in calculating the median. In case of the even numbers the average of the two middle ranking values will be the median.

Mode

Mode is the maximum occurrence or frequency at a particular point or value. You may notice that each one of these measures is a different method of determining a single representative number suited to different types of the data sets.

Mean

Mean is the simple arithmetic average of the different values of a variable. For ungrouped and grouped data, the methods for calculating mean are necessarily different. Mean can be calculated by direct or indirect methods, for both grouped and ungrouped data.

Computing Mean from Ungrouped Data

Direct Method

While calculating mean from ungrouped data using the direct method, the values for each observation are added and the total number of occurrences are divided by the sum of all observations. The mean is calculated using the following formula:

$$\bar{X} = \frac{\sum x}{N}$$

Where,

$\bar{X}$ = Mean

$\sum$ = Sum of a series of measures

x = A raw score in a series of measures

$\sum x$ = The sum of all the measures

N = Number of measures

Example 2.1 : Calculate the mean rainfall for Malwa Plateau in Madhya Pradesh from the rainfall of the districts of the region given in *Table 2.1:*

Table 2.1 : Calculation of Mean Rainfall

Districts in Malwa Plateau	*Normal Rainfall in mms*	*Indirect Method*
	x Direct Method	***d= x-800****
Indore	979	179
Dewas	1083	283
Dhar	833	33
Ratlam	896	96
Ujjain	891	91
Mandsaur	825	25
Shajapur	977	177
$\sum x$ and $\sum d$	**6484**	**884**
$\frac{\sum x}{N}$ and $\frac{\sum d}{N}$	**926.29**	**126.29**

* Where 800 is assumed mean.
d is deviation from the assumed mean.

The mean for the data given in *Table 2.1* is computed as under:

$$\overline{X} = \frac{\sum x}{N}$$

$$= \frac{6,484}{7}$$

$$= 926.29$$

It could be noted from the computation of the mean that the raw rainfall data have been added directly and the sum is divided by the number of observations i. e., districts. Therefore, it is known as **direct method**.

Indirect Method

For a large number of observations, the indirect method is normally used to compute the mean. It helps in reducing the values of the observations to smaller numbers by subtracting a constant value from them. For example, as shown in *Table 2.1*, the rainfall values lie between 800 and 1100 mm. We can reduce these values by selecting 'assumed mean' and subtracting the chosen number from each value. In the present case, we have taken 800 as assumed mean. Such an operation is known as **coding**. The mean is then worked out from these reduced numbers (Column 3 of *Table 2.1*).

The following formula is used in computing the mean using indirect method:

$$\overline{X} = A + \frac{\sum d}{N}$$

Where,

A = Subtracted constant

$\sum d$ = Sum of the coded scores

N = Number of individual observations in a series

Mean for the data as shown in *Table 2.1* can be computed using the indirect method in the following manner :

$$\overline{X} = 800 + \frac{884}{7}$$

$$= 800 + \frac{884}{7}$$

$$\overline{X} = 926.29 \text{ mm}$$

Note that the mean value comes the same when computed either of the two methods.

Computing Mean from Grouped Data

The mean is also computed for the grouped data using either direct or indirect method.

Direct Method

When scores are grouped into a frequency distribution, the individual values lose their identity. These values are represented by the midpoints of the class

intervals in which they are located. While computing the mean from grouped data using direct method, the midpoint of each class interval is multiplied with its corresponding frequency (f); all values of fx (the X are the midpoints) are added to obtain $\sum fx$ that is finally divided by the number of observations i. e., N. Hence, mean is calculated using the following formula :

$$\bar{X} = \frac{\sum fx}{N}$$

Where :

$\bar{X}$ = Mean
f = Frequencies
x = Midpoints of class intervals
N = Number of observations (it may also be defined as $\sum f$)

Example 2.2 : Compute the average wage rate of factory workers using data given in *Table 2.2*:

Table 2.2 : Wage Rate of Factory Workers

Wage Rate (Rs./day)	***Number of workers (f)***
Classes	***f***
50 - 70	10
70 - 90	20
90 - 110	25
110 - 130	35
130 - 150	9

Table 2.3 : Computation of Mean

Classes	***Frequency (f)***	***Mid-points (x)***	***fx***	***d=x-100***	***fd***	***U = (x-100)/ 20***	***fu***
50-70	10	60	600	-40	-400	-2	-20
70-90	20	80	1,600	-20	-400	-1	-20
90-110	**25**	**100**	2,500	**0**	0	0	0
110-130	35	120	4,200	20	700	1	35
130-150	9	140	1,260	40	360	2	18
$\sum fx$ and $\sum fx$	$\sum f$ =99		$\sum fx$ = 10,160		$\sum fd$ = 260		$\sum fu$ = 13

Where N = $\sum f$ = 99

Table 2.3 provides the procedure for calculating the mean for grouped data. In the given frequency distribution, ninety-nine workers have been grouped into five classes of wage rates. The midpoints of these groups are listed in the third column. To find the mean, each midpoint (X) has been multiplied by the frequency (f) and their sum ($\sum fx$) divided by N.

The mean may be computed as under using the given formula :

$$\bar{X} = \frac{\sum fx}{N}$$

$$= \frac{10,160}{99}$$

$$= 102.6$$

Indirect Method

The following formula can be used for the indirect method for grouped data. The principles of this formula are similar to that of the indirect method given for ungrouped data. It is expressed as under

$$\bar{x} = A \pm \frac{\sum fd}{N}$$

Where,

A = Midpoint of the assumed mean group
(The assumed mean group in *Table 2.3* is 90 – 110 with 100 as midpoint.)

f = Frequency

d = Deviation from the assumed mean group (A)

N = Sum of cases or $\sum f$

i = Interval width (in this case, it is 20)

From *Table 2.3* the following steps involved in computing mean using the direct method can be deduced :

(i) Mean has been assumed in the group of 90 – 110. It is preferably assumed from the class as near to the middle of the series as possible. This procedure minimises the magnitude of computation. In *Table 2.3*, A (assumed mean) is 100, the midpoint of the class 90 – 110.

(ii) The fifth column (u) lists the deviations of midpoint of each class from the midpoint of the assumed mean group (90 – 110).

(iii) The sixth column shows the multiplied values of each f by its corresponding d to give fd. Then, positive and negative values of fd are added separately and their absolute difference is found ($\sum fd$). Note that the sign attached to $\sum fd$ is replaced in the formula following A, where ± is given.

The mean using indirect method is computed as under :

$$\bar{x} = A \pm \frac{\sum fd}{N}$$

$$= 100 + \frac{260}{99}$$

$$= 100 + 2.6$$

$$= 102.6$$

Note : The Indirect mean method will work for both equal and unequal class intervals.

Median

Median is a **positional average**. It may be defined "as the point in a distribution with an equal number of cases on each side of it". The **Median** is expressed using symbol M.

Computing Median for Ungrouped Data

When the scores are ungrouped, these are arranged in ascending or descending order. Median can be found by locating the central observation or value in the arranged series. The central value may be located from either end of the series arranged in ascending or descending order. The following equation is used to compute the median :

$$\text{Value of } \left(\frac{N+1}{2}\right) \text{ th item}$$

Example 2.3: Calculate median height of mountain peaks in parts of the Himalayas using the following:

8,126 m, 8,611m, 7,817 m, 8,172 m, 8,076 m, 8,848 m, 8,598 m.

Computation : Median (M) may be calculated in the following steps :

(i) Arrange the given data in ascending or descending order.

(ii) Apply the formula for locating the central value in the series. Thus :

$$\text{Value of } \left(\frac{N+1}{2}\right) \text{ th item}$$

$$= \left(\frac{7+1}{2}\right) \text{th item}$$

$$= \left(\frac{8}{2}\right) \text{th item}$$

4th item in the arranged series will be the Median.

Arrangement of data in ascending order –

7,817; 8,076; 8,126; 8,172; 8,598; 8,611; 8,848

↓

4th item

Hence,

M = 8,172 m

Computing Median for Grouped Data

When the scores are grouped, we have to find the value of the point where an individual or observation is centrally located in the group. It can be computed using the following formula :

$$M = l + \frac{i}{f}\left(\frac{N}{2} - c\right)$$

Where,

M = Median for grouped data
l = Lower limit of the median class
i = Interval
f = Frequency of the median class
N = Total number of frequencies or number of observations
c = Cumulative frequency of the pre-median class.

Example 2.4 : Calculate the median for the following distribution :

class	50-60	60-70	70-80	80-90	90-100	100-110
f	3	7	11	16	8	5

Table 2.4 : Computation of Median

Class	*Frequency (f)*	*Cumulative Frequency (F)*	*Calculation of Median Class*
50-60	3	3	
60-70	7	10	
70-80	11	*21c*	
80-90 (median group)	***16 f***	**37**	$M=\frac{N}{2}$
90-100	8	45	
100-110	5	50	$=\frac{50}{2}$
	$\sum f$ **or N= 50**		$= 25$

The median is computed in the steps given below :

(i) The frequency table is set up as in *Table 2.4*.

(ii) Cumulative frequencies (**F**) are obtained by adding each normal frequency of the successive interval groups, as given in column 3 of *Table 2.4*.

(iii) Median number is obtained by $\frac{N}{2}$ i.e. $\frac{50}{2}$ = **25** in this case, as shown in column 4 of *Table 2.4*.

(iv) Count into the cumulative frequency distribution (**F**) from the top towards bottom until the value next greater than $\frac{N}{2}$ is reached. In this example, $\frac{N}{2}$ is 25, which falls in the Class interval of 40-44 with cumulative frequency of 37, thus the cumulative frequency of the pre-median class is 21 and actual frequency of the median class is 16.

(v) The median is then computed by substituting all the values determined in the step 4 in the following equation :

$$M = l + \frac{i}{f}(m - c)$$

$$= 80 + \frac{10}{16}(25 - 21)$$

$$= 80 + \frac{5}{8} \times 4$$

$$= 80 + \frac{5}{2}$$

$$= 80 + 2.5$$

$$M = 82.5$$

Mode

The value that occurs most frequently in a distribution is referred to as **mode**. It is symbolised as **Z** or $\mathbf{M_0}$. Mode is a measure that is less widely used compared to mean and median. There can be more than one type mode in a given data set.

Computing Mode for Ungrouped Data

While computing mode from the given data sets all measures are first arranged in ascending or descending order. It helps in identifying the most frequently occurring measure easily.

Example 2.5 : Calculate mode for the following test scores in geography for ten students :

61, 10, 88, 37, 61, 72, 55, 61, 46, 22

Computation : To find the mode the measures are arranged in ascending order as given below:

10, 22, 37, 46, 55, **61, 61, 61**, 72, 88.

The measure 61 occurring three times in the series is the **mode** in the given dataset. As no other number is in the similar way in the dataset, it possesses the property of being **unimodal**.

Example 2.6 : Calculate the mode using a different sample of ten other students, who scored:

82, 11, 57, 82, 08, 11, 82, 95, 41, 11.

Computation : Arrange the given measures in an ascending order as shown below :

08, 11, 11, 11, 41, 57, 82, 82, 82, 95

It can easily be observed that measures of 11 and 82 both are occurring three times in the distribution. The dataset, therefore, is **bimodal** in appearance. If three values have equal and highest frequency, the series is **trimodal**. Similarly, a recurrence of many measures in a series makes it **multimodal**. However, when there is no measure being repeated in a series it is designated as **without mode**.

Comparison of Mean, Median and Mode

The three measures of the **central tendency** could easily be compared with the help of normal distribution curve. The normal curve refers to a frequency distribution in which the graph of scores often called a bell-shaped curve. Many

human traits such as intelligence, personality scores and student achievements have normal distributions. The bell-shaped curve looks the way it does, as it is symmetrical. In other words, most of the observations lie on and around the middle value. As one approaches the extreme values, the number of observations reduces in a symmetrical manner. A normal curve can have high or low data variability. An example of a normal distribution curve is given in *Fig. 2.3*.

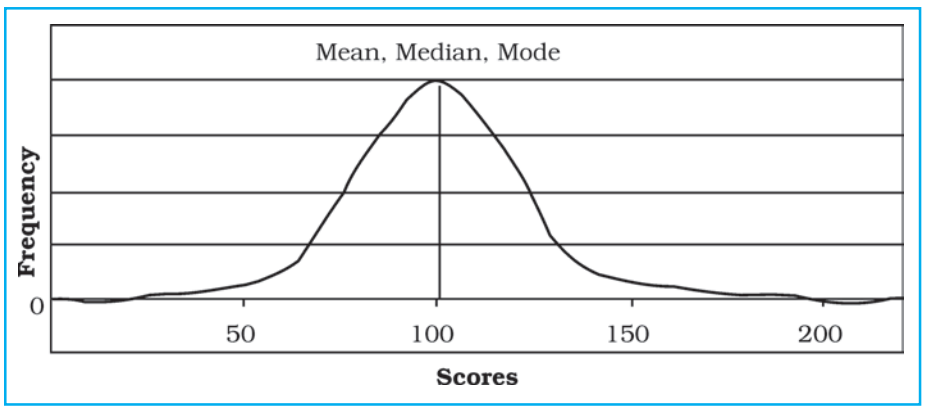

Fig. 2.3 : *Normal Distribution Curve*

The normal distribution has an important characteristic. **The mean, median and mode are the same score** (a score of 100 in *Fig. 2.3*) because a normal distribution is symmetrical. The score with the highest frequency occurs in the middle of the distribution and exactly half of the scores occur above the middle and half of the scores occur below. Most of the scores occur around the middle of the distribution or the mean. Very high and very low scores do not occur frequently and are, therefore, considered rare.

If the data are skewed or distorted in some way, the mean, median and mode will not coincide and the effect of the skewed data needs to be considered (*Fig. 2.4 and 2.5*).

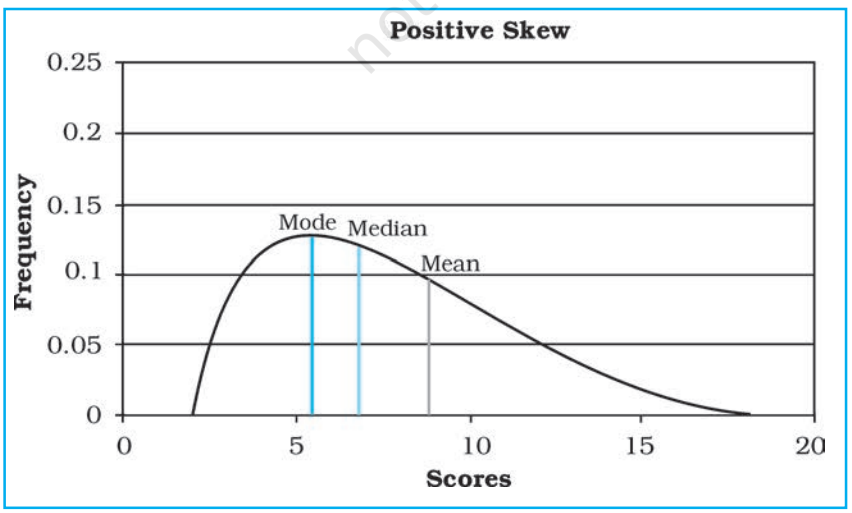

Fig. 2.4 : *Positive Skew*

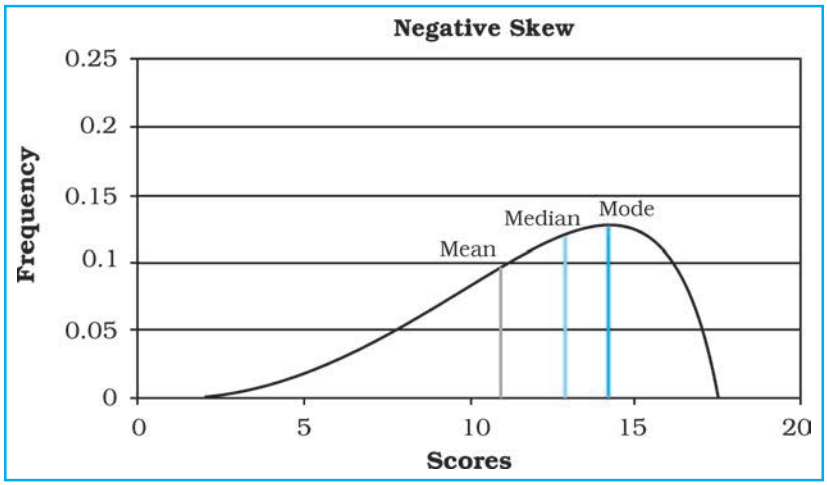

Fig. 2.5 : Negative Skew

Measures of Dispersion

The measures of Central tendency alone do not adequately describe a distribution as they simply locate the centre of a distribution and do not tell us anything about how the scores or measurements are scattered in relation to the centre. Let us use the data given in *Table 2.5* and *2.6* to understand the limitations of the measures of central tendency.

Table 2.5 : Scores of Individuals

Individual	Score
X1	52
X2	55
X3	50
X4	48
X5	45

Table 2.6 : Scores of Individuals

Individual	Score
X1	28
X2	00
X3	98
X4	55
X5	69

$\overline{X}$ = 50 for both the distributions

It can be observed that the mean derived from the two data sets (*Table 2.5* and *2.6*) is same i. e. 50. The highest and the lowest score shown in *Table 2.5* is 55 and 45 respectively. The distribution in *Table 2.6* has a high score of 98 and a low score of zero. The **range** of the first distribution is 10, whereas, it is 98 in the second distribution. Although, the mean for both the groups is the same, the first group is obviously **stable or homogeneous** as compared to the distribution of score of the second group, which is highly **unstable or heterogeneous**. This raises a question whether the mean is a sufficient indicator of the total character of distributions. The examples provide profound evidence that it is not so. Thus, to get a better picture of a distribution, we need to use a measure of **central tendency** and of **dispersion** or **variability.**

The term **dispersion** refers to the scattering of scores about the measure of central tendency. It is used to measure the extent to which individual items or numerical data tend to vary or spread about an average value. Thus, the

dispersion is the degree of spread or scatter or variation of measures about a central value.

The dispersion serves the following two basic purposes :

(i) It gives us the nature of composition of a series or distribution, and

(ii) It permits comparison of the given distributions in terms of stability or homogeneity.

Methods of Measuring Dispersion

The following methods are used as measures of dispersion :

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation and Coefficient of Variation (CV)
5. Lorenz Curve

Each of these methods has definite advantages as well as limitations. Hence, there is a need to use either of the methods with great precautions. The Standard Deviation (s) as an absolute measure of dispersion and Coefficient of Variation (**CV**) as a relative measure of dispersion, besides the Range are most commonly used measures of dispersion. We will discuss how each one of these measures is computed.

Range

Range (R) is the difference between maximum and minimum values in a series of distribution. This way it simply represents the distance from the smallest to the largest score in a series. It can also be defined as the highest score minus the lowest score.

Range for Ungrouped Data

Example 2.7 : Calculate the **range** for the following distribution of daily wages:

Rs. 40, 42, 45, 48, 50, 52, 55, 58, 60, 100.

Computation of Range

The **R** can be calculated with the help of the following formula :

$$R = L - S$$

Where

'R' is Range,

'L' and 'S' is the largest and smallest values respectively in a series.

Hence,

R = L – S

= 100 – 40 = **60**

If we eliminate the 10th case, **R** becomes 20 (60 – 40). The elimination of one score has reduced the **R** to just one-third. It is obvious that the difficulty with **R** as a measure of variability is that its value is wholly dependent upon the two extreme scores. Thus, as a measure of dispersion **R** functions much the same way as **mode** does as a measure of central tendency. Both the measures are **highly unstable**.

Standard Deviation

Standard deviation (SD) is the most widely used measure of dispersion. It is defined as the square root of the average of squares of deviations. It is always calculated around the mean. The standard deviation is the most stable measure of variability and is used in so many other statistical operations. The Greek character σ denotes it.

To obtain SD, deviation of each score from the mean **(*x*)** is first squared (**x^2**). It is important to note that this step makes all negative signs of deviations positive. It saves SD from the major criticism of mean deviation which uses modulus ***x***. Then, all of the squared deviations are summed - **x^2** (care should be taken that these are **not** summed first and then squared). This sum of the squared deviations **(x^2)** is divided by the number of cases and then the square root is taken. Therefore, **Standard Deviation is defined as the root mean square deviation**. For a given data set, it is computed using the following formula :

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

During these steps, we come across a term before taking its square root. It is assigned a special name, the **variance**. The variance is widely used in advanced statistical operations. Its square root is standard deviation. That way, the opposite is also true i.e. square of SD is variance.

Standard Deviation for Ungrouped Data

Example 2.8 : Calculate the standard deviation for the following scores :

01, 03, 05, 07, 09

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

$$= \sqrt{\frac{40}{5}}$$

$= \sqrt{8} = 2.828$ $\sqrt{8}$ 2.828

2.83

Table 2.7 : Computation of Standard Deviation

X	$x(X - \overline{X})$	x^2
1	–4	16
3	–2/–6	4
5	0	0
7	2	4
9	6-Apr	16
$\sum X = 25$		
N =5		
$\therefore = 5$		

Let us summarise the steps used in the above computation :

(i) All the scores have been placed in the column marked **X**.

(ii) Summing the raw scores and dividing by N have found mean.

(iii) Deviation of each raw score **(*x*)** has been obtained by subtracting the mean from them. A check on our work is that the sum of the ***x*** should be zero. We find that this is true for our exercise.

(iv) Each value of ***x*** has been squared and summed.

(v) Sum of the **x^2s** has been divided by N. Recall that the resultant is the variance.

(vi) Its square root has been found to obtain Standard Deviation.

Computation of Standard Deviation for Grouped Data

Example : Calculate the standard deviation for the following distribution:

Groups	120-130	130-140	140-150	150-160	160-170	170-180
f	2	4	6	12	10	6

The method of obtaining SD for grouped data has been explained in the table below. The initial steps upto column 4, are the same as those we followed in the computation of the mean for grouped data. We begin with assuming our mean to exist in the interval group of 150-160, hence a deviation value of zero has been assigned to the group. Likewise other deviations are determined. Values in column 4 (fx') are obtained by the multiplication of the values in the two previous columns. Values in column 5 ($fx'2$) are obtained by multiplying the values given in column 3 and 4. Then various columns have been summed.

(1) ***Group***	(2) ***f***	(3) ***x'***	(4) ***fx'***	(5) ***fx'2***
120 - 130	2	–3	–6	18
130 - 140	4	–2	–8	16
140 - 150	6	–1	$\frac{-6}{-20}$	6
150 - 160	12	0	0	0
160 - 170	10	1	10	10
170 - 180	6	2	$\frac{12}{22}$	24
	N=40		$\sum fx' = 2$	$\sum fx'2 = 74$

The following formula is used to calculate the Standard Deviation :

$$SD = i^2 \left| \sum \mathrm{fx}'^2 - \frac{\sum fx'}{N} \right.$$

Coefficient of Variation (CV)

When the observations for different places or periods are expressed in different units of measurement and are to be compared, the coefficient of variation (CV) proves very useful. **CV expresses the standard deviation as a percentage of the mean.** It is determined using the following formula :

$$\mathrm{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

The CV for the dataset given in *Table 2.7* will, hence, be as under :

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

$$CV = \frac{2.83}{5} \times 100$$

CV = **56%**

Coefficient of Variation for grouped data can also be calculated using the same formula.

Rank Correlation

The statistical methods discussed so far were concerned with the analysis of a single variable. We will now discuss the methods of exploring relationship between two variables and the way this relationship is expressed numerically. When dealing with two or more sets of data, curiosity arises for knowing whether or not changes in one variable produce changes in some other variable.

Often our interest lies in knowing the nature of relationship or interdependence between two or more sets of data. It has been found that the **correlation** serves useful purpose. It is basically a measure of relationship between two or more sets of data. Since, we study the way they vary, we call these events **variables**. Thus, the term **correlation refers to the nature and strength of correspondence or relationship between two variables**. The terms **nature** and **strength** in the definition refer to the **direction** and **degree** of the variables with which they co-vary.

Direction of Correlation

It is our common experience that an input is made to get some output. There could be three possibilities.

1. With the increase in input the output also increases.
2. With the increase in the input the output decreases.
3. Change in the input does not lead to change in the output.

In the first case, the direction of the relationship between the input and output is in the same direction. It is called that both are positively correlated.

In the second case the direction of change between the input and output is in the opposite direction and it is called that they are negatively correlated.

In the third case, change in the input has no relationship with the output, hence, it is said that these do not have a statistically significant relationship.

Let us now consider *Fig. 2.7* which looks just opposite of *Fig. 2.6*. The plotted values run from the upper left to the lower right of the graph. Notice that for every increase of one unit on the X-axis, there is a corresponding decrease of two units on the Y-axis. It is an example of a **negative correlation**. It means that the two variables have a tendency to move opposite to each other, i.e. if one variable increases, the other decreases and vice versa. We can find such relationships existing between various geographical pairs of variables. Correlations between

height above sea level and air pressure, temperature and air pressure are a few examples. It implies that the obtained figure of correlation must precede with the arithmetical sign (plus or minus), more importantly in the negative correlation.

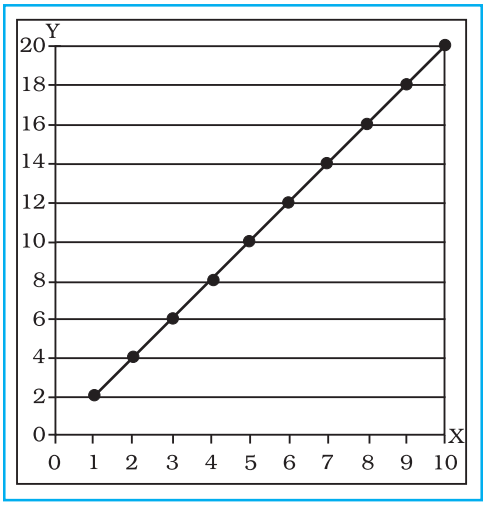

Fig. 2.6 : *Perfect Positive Correlation*

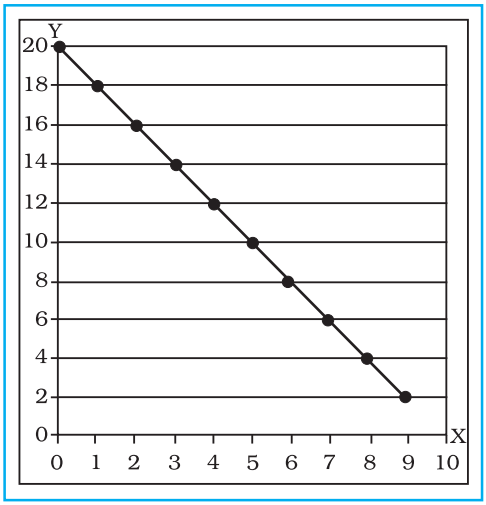

Fig. 2.7 : *Perfect Negative Correlation*

Degree of Correlation

When reference has been made about the direction of correlation, negative or positive, a natural curiosity arises to know the degree of correspondence or association of the two variables. The maximum degree of correspondence or relationship goes upto **1 (one)** in mathematical terms. On adding an element of the direction of correlation, it spreads to the **maximum extent of –1 to +1 through zero**. It can never be more than one. The spread can also be translated into linear shape, as shown in the *Fig. 2.8*. Correlation of **1** is known as **perfect correlation** (whether positive and negative). Between the two points of divergent, perfect correlations lies **0 (zero) correlation,** a point of **no correlation** or absence of any correlation between the variables.

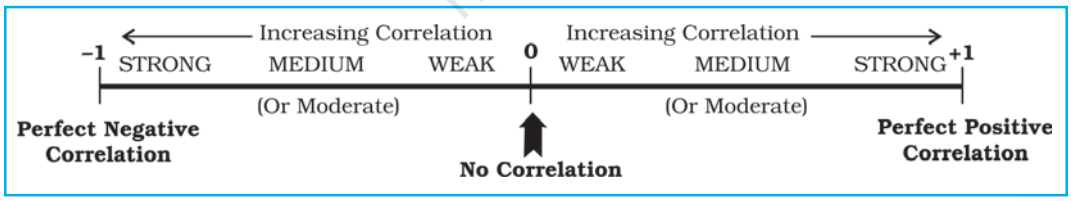

Fig. 2.8 : *Spread of Direction and Degree of Correlation*

Perfect Correlations

Figs. 2.6 and *2.7* have been constructed to show the typical relationship between two variables. Notice that these graphs show the scattering of X – Y values. Therefore, such graphs are referred to as **scatter gram** or **scatter plot**. It may be noted from *Fig. 2.6*, that the pairs of values like these, when plotted, fall along a straight line and when this straight line runs from the lower left of the scatter plot to the upper right, it is an example of a **perfect positive correlation (1.00)**.

Fig. 2.7 is just opposite of this. All the points again fall along a straight line which now runs from the upper left-hand part of the scatter gram to its lower right. It is an example of a **perfect negative correlation** (with a value of – 1.00). **No Correlation** (or Zero Correlation) is one when any of the variables in the pair does not respond to the changes in the other, the correlation will come to zero. This is the state of **no correlation or zero correlation**. This is shown in *Fig. 2.9*. Scatter plot A shows no correlation when Y does not respond to changes in X. Similarly, zero correlation occurs in Seatter plot B when X does not respond to changes in Y.

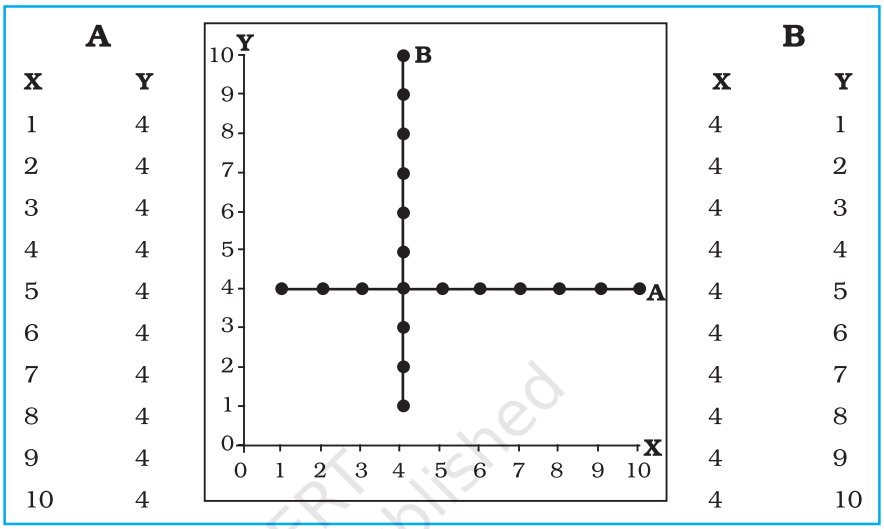

Fig. 2.9 : *Scatter plot showing No Correlation*

Other Correlations

Between the perfect correlations (±1) and zero correlation lies generalised conditions popularly referred to as weak, moderate and strong correlations. These conditions are clearly exhibited in *Figs. 2.10, 2.11* and *2.12* respectively. Notice the spreading or the scattering of the plotted points and the assignment of the terms weak, medium and strong to them (generalised terms having no specific limits). Larger is the scattering, weaker is the correlation. Smaller is the scattering, stronger is the correlation, and when the plotted points fall on a straight line, the correlation is perfect (*Fig. 2.6 and 2.7*).

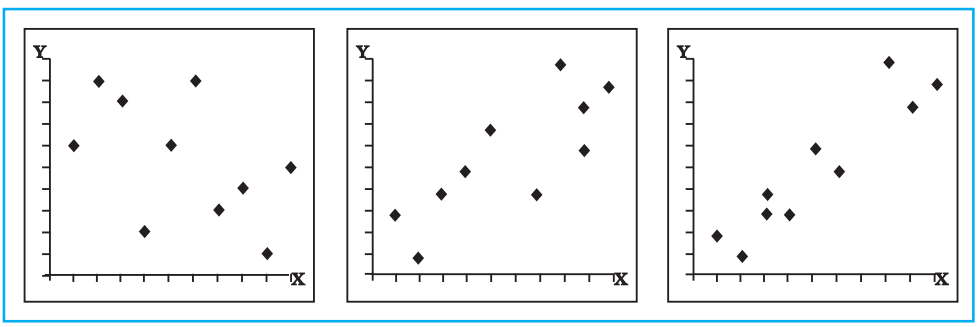

Fig. 2.10 : *Weak Negative Correlation*

Fig. 2.11 : *Moderate Positive Correlation*

Fig. 2.12 : *Strong Positive Correlation*

Methods of Calculating Correlation

There are various methods by which correlation can be calculated. However, under the constraints of time and space, we will discuss the Spearman's Rank Correlation method only.

Spearman's Rank Correlation

Spearman devised a method of computing correlation with the help of ranks. The method is popularly known as Spearman's Rank Correlation symbolised asr ρ (the Greek letter **rho**). Spearman's Rank Correlation method is widely used. The computation of the correlation is undertaken in the steps given below:

(i) Copy the data related to X-Y variables given in the exercise and put them in the first and second columns of the table.

(ii) Both the variables are to be ranked separately. The ranks of X-variable are to be recorded in column 3 headed by **XR** (ranks of X). Similarly, the ranks of Y-variable **(YR)** are to be recorded in the fourth column. The highest value in the data is to be awarded rank one, second highest rank two and so on. Suppose the data for X-variable are 4, 8, 2, 10, 1, 9, 7, 3, 0 and 5, the **XR** will be 6, 3, 8, 1, 9, 2, 4, 7, 10 and 5 respectively. Notice that the last rank (10 in this case) equals the number of observations. Assignment of **YR** is also done in the same way.

(iii) Now since both **XR** and **YR** have been obtained, find the difference between the two sets of ranks (disregarding the sign plus or minus) and record it in the fifth column. **The sign of the difference is of no importance**, since, these differences are squared in the next operation.

(iv) Each of these differences is squared and sum of this column of squares is obtained. These values are placed in the sixth column.

(v) Then the computation of the rank correlation is done by the application of the following equation:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

Where,

ρ = rank correlation

$\sum D^2$ = sum of the squares of the differences between two sets of ranks

N = the number of pairs of X-Y

Example 2.9: Calculate Spearman's Rank Correlation with the help of the following data :

Scores in Economics (X) :	02	08	00	20	12	16	06	18	09	10
Scores in Geography (Y) :	04	12	06	24	16	18	08	20	09	10

Table 2.8 : Computation of Spearman's Rank Correlation

(1) *X*	(2) *Y*	(3) *XR*	(4) *YR*	(5) *D*	(6) D^2
2	4	9	10	1	1
8	12	7	5	2	4
0	6	10	9	1	1
20	24	1	1	0	0
12	16	4	4	0	0
16	18	3	3	0	0
6	8	8	8	0	0
18	20	2	2	0	0
9	9	6	7	1	1
10	10	5	6	1	1
N=10					D^2=8

Calculation:

Where, ρ is Rank Correlation; **D** is difference between the rank of X and Y; and **N** is number of items of x – y

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \times 8}{10\left(10^2 - 1\right)}$$

$$= 1 - \frac{48}{10\left(100 - 1\right)}$$

$$= 1 - \frac{48}{10\left(99\right)}$$

$$= 1 - \frac{48}{\left(990\right)}$$

$$= 1 - 0.05$$

$$= 0.95$$

In rho, we obtain a correlation, which makes a good substitute for other types of correlations, when the number of cases is small. It is almost useless when N is large, because by the time all the data are ranked, other type of correlation could have been calculated.

Excercises

1. Choose the correct answer from the four alternatives given below:

   (i) The measure of central tendency that does not get affected by extreme values:
   (a) Mean (b) Mean and Mode
   (c) Mode (d) Median

   (ii) The measure of central tendency always coinciding with the hump of any distribution is:
   (a) Median (b) Median and Mode
   (c) Mean (d) Mode

   (iii) A scatter plot represents negative correlation if the plotted values run from:
   (a) Upper left to lower right (b) Lower left to upper right
   (c) Left to right (d) Upper right to lower left

2. Answer the following questions in about 30 words:

   (i) Define the mean.
   (ii) What are the advantages of using mode ?
   (iii) What is dispersion ?
   (iv) Define correlation.
   (v) What is perfect correlation ?
   (vi) What is the maximum extent of correlation?

3. Answer the following questions in about 125 words:

   (i) Explain relative positions of mean, median and mode in a normal distribution and skewed distribution with the help of diagrams.
   (ii) Comment on the applicability of mean, median and mode (*hint: from their merits and demerits).*
   (iii) Explain the process of computing Standard Deviation with the help of an imaginary example.
   (iv) Which measure of dispersion is the most unstable statistic and why?
   (v) Write a detailed note on the degree of correlation.
   (vi) What are various steps for the calculation of rank order correlation?

Activity

1. Take an imaginary example applicable to geographical analysis and explain direct and indirect methods of calculating mean from ungrouped data.
2. Draw scatter plots showing different types of perfect correlations.