UNIT 8 MEASURES OF VARIATION AND SKEWNESS

Objectives

After going through this unit, you will learn:

- the concept and significance of measuring variability
- the concept of absolute and relative variation
- the computation of several measures of variation, such as the range, quartile
- deviation, average deviation and standard deviation and also their coefficients
- the concept of skewness and its importance
- the computation of coefficient of skewness.

Structure

8.1 INTRODUCTION

In the previous unit, we were concerned with various measures that are used to provide a single representative value of a given set of data. This single value alone cannot adequately describe a set of data. Therefore, in this unit, we shall study two more important characteristics of a distribution. First we shall discuss the concept of variation and later the concept of skewness.

A measure of variation (or dispersion) describes the spread or scattering of the individual values around the central value. To illustrate the concept of variation, let us consider the data given below:

Firm A Daily Sales (Rs.)	Firm B Daily Sales (Rs.)	Firm C Daily Sales (Rs.)
5000	5050	4900
5000	5025	3100
5000	4950	2200
5000	4835	1800
5000	5140	13000
$\bar{X}_A = 5000$	$\bar{X}_B = 5000$	$\bar{X}_C = 5000$

Since the average sales for firms A, B and C is the same, we are likely to conclude that the distribution pattern of the sales is similar. It may be observed that in Firm A, daily sales are the same irrespective of the day, whereas there is less amount of variation in the daily sales for firm 13 and greater amount of variation in the daily sales for firm C. Therefore, different sets of data may have the same measure central tendency but differ greatly in terms of variation.

8.2 SIGNIFICANCE OF MEASURING VARIATION

Measuring variation is significant for some of the following purposes.

i) Measuring variability determines the reliability of an average by pointing out as to how far an average is representative of the entire data.

ii) Another purpose of measuring variability is to determine the nature and cause variation in order to control the variation itself.

iii) Measures of variation enable comparisons of two or more distributions with regard to their variability.

iv) Measuring variability is of great importance to advanced statistical analysis. For example, sampling or statistical inference is essentially a problem in measuring variability.

8.3 PROPERTIES OF A GOOD MEASURE OF VARIATION

A good measure of variation should possess, as far as possible, the same properties as those of a good measure of central tendency.

Following are some of the well known measures of variation which provide a numerical index of the variability of the given data:

i) Range

ii) Average or Mean Deviation

iii) Quartile Deviation or Semi-Interquartile Range

iv) Standard Deviation

8.4 ABSOLUTE AND RELATIVE MEASURES OF VARIATION

Measures of variation may be either absolute or relative. Measures of absolute variation are expressed in terms of the original data. In case the two sets of data are expressed in different units of measurement, then the absolute measures of variation are not comparable. In such cases, measures of relative variation should be used. The other type of comparison for which measures of relative variation are used involves the comparison between two sets of data having the same unit of measurement but with different means. We shall now consider in turn each of the four measures of variation.

8.5 RANGE

The range is defined as the difference between the highest (numerically largest) value and the lowest (numerically smallest) value in a set of data. In symbols, this may be indicated as:

$R = H - L,$

where R = Range; H = Highest Value; L = Lowest Value

As an illustration, consider the daily sales data for the three firms as given earlier.

For firm A, $R = H - L = 5000 - 5000 = 0$

For firm B, $R = H - L = 5140 - 4835 = 305$

For firm C, $R = H - L = 13000 - 18000 = 11200$

The interpretation for the value of range is very simple.

In this example, the variation is nil in case of daily sales for firm A, the variation is small in case of firm B and variation is very large in case of firm C.

The range is very easy to calculate and it gives us some idea about the variability of the data. However, the range is a crude measure of variation, since it uses only two extreme values.

The concept of range is extensively used in statistical quality control. Range is helpful in studying the variations in the prices of shares and debentures and other commodities that are very sensitive to price changes from one period to another. For meteorological departments, the range is **a** good indicator for weather forecast.

For grouped data, the range may be approximated as the difference between the upper limit of the largest class and the lower limit of the smallest class.

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula

$$\text{Coefficient of range} = \frac{H - L}{H + L}$$

Activity A

Following are the prices of shares of a company from Monday to Friday:

Day	:	Monday	Tuesday	Wednesday	Thursday	Friday
Price	:	670	678	750	705	720

Compute the value of range and interpret the value.

……
……
……
……

Activity B

Calculate the coefficient of range from the following data:

Sales (Rs. lakhs)	No. of days	Sales (Rs. lakhs)	No. of days
30–40	12	60–70	19
40–50	18	70–80	13
50–60	20	80–90	8

……
……
……
……

8.6 QUARTILE DEVIATION

The quartile deviation, also known as semi-interquartile range, is computed by taking the average of the difference between the third quartile and the first quartile. In symbols, this can be written as:

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

where Q_1 = first quartile, and Q_3 = third quartile.

The following illustration would clarify the procedure involved. For the data given below, compute the quartile deviation.

Monthly Wages (Rs.)	No. of workers	Monthly Wages (Rs.)	No. of Workers
Below 850	12	1000–1050	62
850–900	16	1050–1100	75
900–950	39	1100–1150	30
950–1000	56	1150 and above	10

To compute quartile deviation, we need the values of the first quartile and the third quartile which can be obtained from the following table:

Monthly Wages (Rs.)	No. of workers f	C.F.
Below 850	12	12
850-900	16	28
900-950	39	67
950 -1000	56	123
1000-1050	62	185
1050-1100	75	260
I100-1150	30	290
1150 and above	I0	300

$Q_1 = \text{Size of } \frac{N}{4}\text{th observation} = \frac{300}{4} = 75\text{th observation which lies in the class } 950 - 1000.$

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i = 950 + \frac{75 - 67}{56} \times 50$$

$$= 950 + \frac{50}{7} = 950 + 7.14 = 957.14$$

$Q_3 = \text{Size of } \frac{3N}{4}\text{th observation} = \frac{3 \times 300}{4} = 25\text{th observation which lies in the class } 1050 - 1100.$

$$Q_3 = L + \frac{3N/4 - pcf}{f} \times i = 1050 + \frac{225 - 185}{75} \times 50$$

$$= 1050 + \frac{2000}{75} = 1050 + 26.67 = 1076.67$$

$$\text{Q.D.} = \frac{1076.67 - 957.14}{2} = \frac{119.53}{2} = 59.765$$

The relative measure corresponding to quartile deviation, called the coefficient of quartile deviation, is calculated as given below:

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The quartile deviation is superior to the range as it is not based on two extreme values but rather on middle 50% observations. Another advantage of quartile deviation is that it is the only measure of variability which can be used for open-end distribution.

The disadvantage of quartile deviation is that it ignores the first and the last 25% observations.

Activity C

A survey of domestic consumption of electricity gave the following distribution of the units consumed. Compute the quartile deviation and its coefficient.

Number of units	Numberofconsumers	Number of units	Numberofconsumers
Below 200	9	800-1000	45
200-400	18	1000-1200	38
400-600	27	1200-1400	20
600-800	32	1400 & above	11

……
……
……
……
……

8.7 AVERAGE DEVIATION

The measure of average (or mean) deviation is an improvement over the previous two measures in that it considers all observations in the given set of data. This measure is computed as the mean of deviations from the mean or the median. All the deviations are treated as positive regardless of sign. In symbols, this can be represented by:

$$\text{A.D.} = \frac{\sum|X - \bar{X}|}{N} \text{ or } \frac{\sum|X - \text{Median}|}{N}$$

Theoretically speaking, there is an advantage in taking the deviations from median because the sum of the absolute deviations (i.e. ignoring ± signs) from median is minimum. In actual practice, however, arithmetic mean is more popularly used in computation of average deviation.

For grouped data, the formula to be used is given as:

$$\text{A.D.} = \frac{\sum|X - \bar{X}|}{N}$$

As an illustration, consider the following grouped data which relate to the sales of 100 companies.

Sales (Rs. thousand)	No. of days	Sales (Rs. thousand)	No. of days
40–50	10	70–80	30
50–60	15	80–90	12
60–70	25	90–100	8

To compute average deviation, we construct the following table:

Sales (Rs. thousand)	X m.p	No. of days	fX	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$
40–50	45	5	225	26	130
50–60	55	15	825	16	240
60–70	65	25	1625	6	150
70–80	75	30	2250	4	120
80–90	85	20	1700	14	280
90–100	95	5	475	24	120
		N = 100	$\sum fX = 7100$		$\Sigma f\lvert X-\bar{X}\rvert = 1040$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{7100}{100} = 71$$

$$\text{A.D.} = \frac{\Sigma f|X-\bar{X}|}{N} = \frac{1040}{100} = 10.4$$

The relative measure corresponding to the average deviation, called the coefficient of average deviation, is obtained by dividing average deviation by the particular average used in computing the average deviation. Thus, if average deviation has been computed from median, the coefficient of average deviation shall be obtained by dividing the average deviation by the median.

$$\text{Coefficient of A.D.} = \frac{\text{A.D.}}{\text{Median}} \text{ or } \frac{\text{A.D.}}{\text{Mean}}$$

Although the average deviation is a good measure of variability, its use is limited. If one desires only to measure and compare variability among several sets of data, the average deviation may be used.

The major disadvantage of the average deviation is its lack of mathematical properties. This is more true because non-use of signs in its calculations makes it algebraically inconsistent.

Activity D

Calculate the average deviation and coefficient of the average deviation from the following data.

Sales (Rs. thousand)	No. of days	Sales (Rs. thousand)	No. of days
Less than 20	3	Less than 50	23
Less than 30	9	Less than 60	25
Less than 40	20		

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

8.8 STANDARD DEVIATION

The standard deviation is the most widely used and important measure of variation. In computing the average deviation, the signs are ignored. The standard deviation overcomes this problem by squaring the deviations, which makes them all positive. The standard deviation, also known as root mean square deviation, is generally denoted by the lower case Greek letter a (read as sigma). In symbols, this can be expressed as

$$\sigma = \sqrt{\frac{\sum (X - \overline{X})^2}{N}}$$

The square of the standard deviation is called variance. Therefore

$$\text{Variance} = \sigma^2$$

The standard deviation and variance become larger as the cm a within the data becomes greater. More important, it is readily comparable with other standard deviations and the greater the standard deviation, the greater the variability.

For grouped data, the formula is

$$\sigma = \sqrt{\frac{\sum f(X - \overline{X})^2}{N}}$$

The following formulas for standard deviation are mathematically equivalent to the above formula and are often more convenient to use in calculations.

$$\sigma = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\frac{\sum fX^2}{N} - \overline{X}^2}$$

$$= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i \qquad \text{Where } d = \frac{X - A}{i}$$

Remarks: If the data represent a sample of size N from a population, then it can be proved that the sum of the squared deviations are divided by (N-1) instead of by N. However, for large sample sizes, there is very little difference in the use of (N-1) or N in computing the standard deviation.

To understand the formula for grouped data, consider the following data which relate to the profits of 100 companies.

Profit (Rs. lakhs)	No. of companies	Profit (Rs. lakhs)	No. of companies
8-10	8	14-16	30
10-12	12	16-18	20
12-14	20	18-20	10

To compute standard deviation we construct the following table:

Profits (Rs. lakhs)	m.p. X	f	d = (X–15)/2	fd	fd^2
8–10	9	8	–3	–24	72
10–12	11	12	–2	–24	48
12–14	13	20	–1	–20	20
14–16	15	30	0	0	0
16–18	17	20	+1	+20	20
18–20	19	10	+2	+20	40
		N = 100		$\sum fd = -28$	$\sum fd^2 = 200$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i = \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 2$$

$$= \sqrt{2 - 0.0784} \times 2 = \sqrt{1.9216} \times 2$$

$$= 1.3862 \times 2 = 2.7724 \simeq 2.77$$

The standard deviation is commonly used to measure variability, while all other measures have rather special uses. In addition, it is the only measure possessing the necessary mathematical properties to make it useful for advanced statistical work.

Activity E

The following data show the daily sales at a petrol station. Calculate the mean and standard deviation.

Number of litres sold	No. of days	Number of litres sold	No. of days
700-1000	12	1900-2200	18
1000-1300	18	2200-2500	5
1300-1600	20	2500-2800	2
1600-1900			

……
……
……
……
……
……
……
……
……
……

8.9 COEFFICENT OF VARIATION

A frequently used relative measure of variation is the coefficient of variation, denoted by C.V. This measure is simply the ratio of the standard deviation to mean expressed as the percentage.

Coefficient of variation = C.V. = $\frac{\sigma}{\overline{X}}$ 100 when the coefficient of variation is less in the data, it is said to be less variable or more consistent.

Consider the following data which relate to the mean daily sales and standard deviation for four regions.

Region	Mean daily sales (Rs. thousand)	Standard deviation (Rs. thousand)
1	86	10.45
2	45	5.86
3	72	9.54
4	61	11.32

To determine which region is most consistent in terms of daily sales, we shall compute the coefficients of variation. You may notice that the mean daily sales are not equal for each region.

$$C.V._1 = \frac{10.45}{86} \times 100 = 12.15;\ C.V._2 = \frac{5.86}{45} \times 100 = 13.02$$

$$C.V._3 = \frac{9.54}{72} \times 100 = 13.25;\ C.V._4 = \frac{11.32}{61} \times 100 = 18.56$$

As the coefficient of variation is minimum for Region1, therefore the most consistent region is Region1.

Activity F

A factory produces two types of electric lamps, A and B. In an experiment re1ating to their life, the following results were obtained.

Length of life (in hours)	Type A No. of lamps	Type B No. of lamps
500-700	5	4
700-900	11	30
900-1100	26	12
1100-1300	10	8
1300-1500	8	6

Compare the variability of the life of the two types of electric lamps using the coefficient of variation.

……
……
……
……
……
……

8.10 SKEWNESS

The measures of central tendency and variation do not reveal all the characteristics of a given set of data. For example, two distributions may have the same mean and

standard deviation but may differ widely in the shape of their distribution. Either the distribution of data is symmetrical or it is not. If the distribution of data is not symmetrical, it is called *asymmetrical* or *skewed.* Thus skewness refers to the lack of symmetry in distribution.

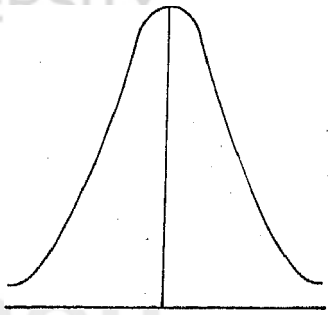
A simple method of detecting the direction of skewness is to consider the tails of the distribution (Figure I). The rules are:

Data are symmetrical when there are no extreme values in a particular direction so that low and high values balance each other. In this case, mean = median = mode. (see Fig I(a)).

If the longer tail is towards the lower value or left hand side, the skewness is negative. Negative skewness arises when the mean is decreased by some extremely low values, thus making mean < median < mode. (see Fig I(b)).

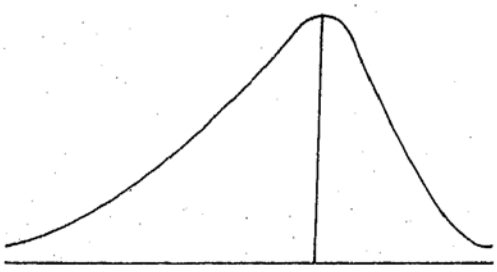
If the longer tail of the distribution is towards the higher values or right hand side, the skewness is positive. Positive skewness occurs when mean is increased by some unusually high values, thereby making mean > median > mode. (see Fig I(c))

Figure I



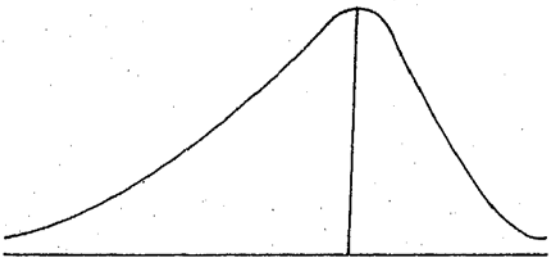
(a)

Symmetrical Distribution



(b)

Negatively skewed Distribution



(c)

Positively skewed distribution

8.11 RELATIVE SKEWNESS

In order to make comparisons between the skewness in two or more distributions, the coefficient of skewness (given by Karl Pearson) can be defined as:

$$SK. = \frac{\text{Mean - Mode}}{\text{S.D.}}$$

If the mode cannot he determined, then using the approximate relationship, Mode = 3 Median - 2 Mean, the above formula reduces to

$$SK. = \frac{3\ (\text{Mean - Median})}{\text{S.D.}}$$

if the value of this coefficient is zero, the distribution is symmetrical; if the value of the coefficient is positive, it is positively skewed distribution, or if the value of the coefficient is negative, it is negatively skewed distribution. In practice, the value of this coefficient usually lies between ± I.

When we are given open-end distributions where extreme values are present in the data or positional measures such as median and quartiles, the following formula for coefficient of skewness (given by Bowley) is more appropriate.

$$SK. = \frac{Q_3 + Q_1 - 2\ \text{Median}}{Q_3 - Q_1}$$

Again if the value of this coefficient is zero, it is a symmetrical distribution. For positive value, it is positively skewed distribution and for negative value, it is negatively skewed distribution.

To explain the concept of coefficient of skewness, let us consider the following data.

Profits (Rs. thousand)	No. of companies	Profits (Rs. thousand)	No. of companies
10-12	7	18-20	25
12-14	15	20-22	10
14-16	18	22-24	5
16-18	20		

Since the given distribution is not open-ended and also the mode can be determined, it is appropriate to apply Karl Pearson formula as given below:

$$SK. = \frac{\text{Mean - Mode}}{\text{S.D.}}$$

Profits (Rs. thousand)	m.p. X	f	d=(X-17)/2	fd	fd^2
10-12	11	7	-3	-21	63
12-14	13	15	-2	-30	60
14-16	15	18	-1	-18	18
16-18	17	20	0	0	0
I8-20	19	25	+1	25	25
20-22	21	10	+2	20	40
22-24	23	5	+3	15	45
		N = 100		$\sum fd = -9$	$\sum fd^2 = 251$

$$\bar{X} = A + \frac{\sum fd}{N} \times i = 17 - \frac{9}{100} \times 2 = 17 - 0.18 = 16.82$$

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2} \times i = 18 + \frac{5}{15} \times 2 = 18 + 0.67 = 18.67$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i = \sqrt{\frac{251}{100} - \left(\frac{-9}{100}\right)^2} \times 2$$

$$\sqrt{2.51 - 0.0081} \times 2 \quad = \sqrt{2.509} \times 2 \quad = 1.5817 \times 2 = 3.1634$$

$$Sk = \frac{16.82 - 18.67}{3.1634} \quad = -0.585$$

This value of coefficient of skewness indicates that the distribution is negatively skewed and hence there is a greater concentration towards the higher profits.

The application of Bowley's method would be clear by considering the following data:

Sales (Rs. lakhs)	No. of companies	c.f.
Below 50	8	8
50-60	12	20
60-70	20	40
70-80	25	65
80 & above	15	80

Q_1 = size of $\frac{N}{4}$th observation $= \frac{80}{4}$ = 20th observation which lies in the class 50–60

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i = 50 + \frac{20 - 8}{12} \times 10 = 60$$

Q_2 = Median = size of $\frac{N}{2}$th observation $= \frac{80}{2}$ = 40th observation which lies in the class 60–70

$$Q_2 = \text{Med.} = L + \frac{N/2 - pcf}{f} \times i = 60 + \frac{40 - 20}{20} \times 10 = 70$$

Q_3 = Size of $\frac{3N}{4}$th observation $= \frac{3 \times 80}{4}$ = 60th observation which lies in the class 70–80

$$Q_3 = L + \frac{3N/4 - pcf}{f} \times i = 70 + \frac{60 - 40}{25} \times 10 = 78$$

$$\text{Coefficient of SK.} = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

$$= \frac{78 + 60 - 2 \times 70}{78 - 60} = -0.11$$

This value of coefficient of skewness indicates that the distribution is slightly skewed to the left and therefore there is a greater concentration of the sales at the higher values than the lower values of the distribution.

8.12 SUMMARY

In this unit, we have shown how the concepts of measures of variation and skewness are important. Measures of variation considered were the range, average deviation,

quartile deviation and standard deviation. The concept of coefficient of variation was used to compare relative variations of different data. The skewness was used in relation to lack of symmetry.

8.13 KEY WORDS

Average Deviation is the arithmetic mean of the absolute deviations from the mean or the median.

Coefficient of Variation is a ratio of standard deviation to mean expressed as percentage.

Interquartile Range considers the spread in the middle 50% ($Q_3 - Q_1$) of the data.

Quartile Deviation is one half the distance between first and third quartiles.

Range is the difference between the largest and the smallest value in a set of data.

Relative Variation is used to compare two or more distributions by relating the variation of one distribution to the variation of the other.

Skewness refers to the lack of symmetry.

Standard Deviation is the root mean square deviation of a given set of data.

Variance is the square of standard deviation and is defined as the arithmetic mean of the squared deviations from the mean.

8.14 SELF- SSESSMENT EXERCISES

1 Discuss the important of measuring variability for managerial decision making.

2 Review the advantages and disadvantages of each of the measures of variation.

3 What is the concept of relative variation? What problem situations call for the use of relative variation in their solution?

4 Distinguish between Karl Pearson's and Bowley's coefficient of skewness. Which one of these would you prefer and why?

5 Compute the range and the quartile deviation for the following data:

Monthly wage (Rs.)	No. of workers	Monthly wage (Rs.)	No. of workers
700-800	28	1000-1100	30
800-900	32	1100-1200	25
900-1000	40	1200-1300	15

6 Compute the average deviation for the following data:

No. of shares applied for	No. of applicants	No. of shares applied for	No. of applicants
50-100	2500	250-300	900
100-150	1500	300-350	750
150-200	1300	350-400	675
200-250	I100	400-450	525
		450-500	450

7 Calculate the mean, standard deviation and variance for the following data

No. of defects per item	Frequency	No. of defects per item	Frequency
0-5	18	25-30	150
5-10	32	30-35	100
10-15	50	35-40	90
15-20	75	40-45	80
20-25	125	45-50	50

8 Records were kept on three employees who wrapped packages on sweet boxes during the Diwali holidays in a big sweet house. The study yielded the following data

Employee	Mean number of packages	Standard deviation
A	23	1.45
B	45	5.86
C	32	3.54

i) Which package wrapper was most productive?
ii) Which employee was the most consistent?
iii) What measure did you choose to answer part (ii) and why?

9 The following data relate to the mileage of two types of tyre:

Life (in kms.)	Number of Tyres	
	Type A	Type B
20000–22000	230	200
22000–24000	270	275
24000–26000	450	470
26000–28000	375	300
28000–30000	125	155

i) Which of the two types gives a higher average life?
ii) If prices are the same for both the types, which would you prefer and why?

10 The following table gives the distribution of daily travelling allowance to salesmen in a company:

Travelling Allowance (in Rs.)	No. of salesmen	Travelling Allowance (in Rs.)	No. of salesmen
100–120	14	180–200	15
120–140	16	200–220	7
140–160	20	220–240	6
160–180	18	240–260	4

Compute Karl Pearson's coefficient of skewness and comment on its value.

11 Calculate Bowley's coefficient of skewness from the following data:

Monthly wages	No. of workers	Monthly wages	No. of workers
Below 600	10	800–900	20
600–700	25	900–1000	15
700–800	45	1000 & above	5

12 You are given the following information before and after the settlement of workers' strike.

	Before settlement of strike	After settlement of strike
No. of workers	1000	950
Average Wage (Rs.)	1300	1350
Standard Deviation (Rs.)	400	425
Median Wage (Rs.)	1325	1300

Assuming that the increase in wage is a loss to the management, comment on the gains and losses from the point of view of workers and that of management.

8.15 FURTHER READINGS

Clark, T.C. and E.W. Jordan, 1985. *Introduction to Business and Economic Statistics,* South-Western Publishing Co.:

Enns, P.G., 1985. *Business Statistics,* Richard D. Irwin Inc.: Homewood.

Gupta, S.P. and M.P. Gupta, 1988. *Business Statistics,* Sultan Chand & Sons: New Delhi.

Moskowitz, H. and G.P. Wright, 1985. *Statistics for Management and Economics,* Charles E. Merill Publishing Company.