UNIT 9 MEASURES OF VARIATION AND SKEWNESS

STRUCTURE

9.0 OBJECTIVES

After studying this Unit, you should be able to:

- describe the concept and significance of measuring variability for data analysis,
- compute various measures of variation and its application for analysing the data,
- choose an appropriate measure of variation under different situations,
- describe the importance of Skewness in data analysis,
- explain and differentiate the symmetrical, positively skewed and negatively skewed data, and
- ascertain the value of the coefficient of skewness and comment on the nature of distribution.

9.1 INTRODUCTION

In Unit 8, we have learnt about the measures of central tendency. They give us only a single figure that represents the entire data. However, central tendency alone cannot adequately describe a set of data, unless all the values of the variables in the collected data are the same. Obviously, no average can sufficiently analyse the data, if all the values in a distribution are widely spread. Hence, the measures of central tendency must be supported and supplemented with other measures for analysing the data more meaningfully. Generally, there are three other characteristics of data which provide useful information for data analysis i.e., Variation, Skewness, and Kurtosis. The third characteristic, Kurtosis, is not with in the scope of this course. In this unit, therefore, we shall discuss the importance of measuring variation and skewness for describing distribution of data and their computation. We shall also discuss the role of normal curves in characterizing the data.

9.2 VARIATION – WHY IS IT IMPORTANT?

Measures of variation are statistics that indicate the degree to which numerical data tend to spread about an average value. It is also called dispersion, scatter, spread etc., It is related to the homogeniety of the data. In the simple words of Simpson of Kafka "the measurement of the scatterness of the mass of figures (data) in a series about an average is called measure of variation". Therefore, we can say, variation measures the extent to which the items scatter from average. To be more specific, an average is more meaningful when the data are examined in the light of variation. Infact, in the absence of measure of dispersion, it will not be possible to say which one of the two or more sets of data is represented more closely and adequately by its arithmetic mean value. Here, the following illustration helps you to understand the necessity of measuring variability of data for effective analysis.

Illustration-1

The data given below relates to the marks secured by three students (A, B and C) in different subjects

Subjects	Marks		
	A	B	C
Research methodology	50	50	10
Accounting for Mangers	50	70	100
Financial Management	50	40	80
Marketing Management	50	40	30
Managerial Economics	50	50	30
Total	250	250	250
Mean ($\bar{x}$)	50	50	50

In the above illustration, you may notice that the marks of the three students have the same mean i.e. the average marks of A, B and C are the same i.e., 50 Marks, and we may analyse and conclude that the three distributions are similar. But, you should note that, by observing distributions (subject-wise) there is a wide difference in the marks of these three students. In case of 'A' the marks in each subject are 50, hence we can say each and every item of the data is perfectly represented by the mean or in other words, there is no variation. In case of B there is slight variation as compared to 'C', where as in case of 'C' not a single item is perfectly represented by the mean and the items vary widely from one another. Thus, different set of data may have the same value of average, but may differ greatly in terms of spread or scatter of items. The study of variability, therefore, is necessary to know the average scatter of the item from the average to gauge the degree of variability in the collected data.

9.3 SIGNIFICANCE OF VARIATION

The measure of variability is useful in various situations. Let us take an example to understand the significance of variation.

A family intends to cross a lake. They come to know that the average depth of the lake is 4 feet. The average height of the family, is 5.5 feet. Then they decide that the lake can be crossed safely. While crossing the lake, at a particular place all the members of the family get drowned where the level of water is more than 6.5 feet deep. The reason for drowning is that they rely on the average depth of the lake and their average height but do not rely on the variability of the Lake's depth and their height. In the light of the above example, we may understand the reason for measuring variability of a given data.

To Judge the reliability of an average: Financial analysts examine the variation of a firm's earnings. If earnings are widely scattered (extremely high to low or even negative) it indicates a high risk to investors. Where there is a wide scatterred in the data, the measure of variation gives a description of the structure of the data. If variation is small, the average closely represents the individual values and may say it is reliable. On the other hand, if the variation is greater the average may be unreliable.

To compare series with regard to their variability: Measuring variation enables us to compare the variability between two or more series. It is useful to study the degree of uniformity and consistency in different data sets. A greater degree of variation in a data set means low degree of consistency. On the other hand, low degree of variation means high degree of consistency in that distribution.

To provide a basis for the control of variability itself: It facilitates to determine the nature and cause of variation in order to control the variation itself. Quality control experts analyse the variation of the quality of a product. For instance, a drug that may be average in purity but varies from very pure to highly impure may endanger lives.

To facilitate the use of other statistical measures: Many analytical devices in statistics such as hypothesis testing, cost control, analysis of fluctuations, correlation and regression analysis, techniques of production control etc. are based on the measure of variation.

Keeping in view the above purposes, the variation of data must be taken into account while taking business decisions.

9.4 MEASURES OF VARIATION

Variation may be measured in absolute or relative terms. Measures of absolute variation are expressed in terms of the original units of the given data. For example, the temperature of a city in a day ranges between 15^0C and 47^0C, then absolute variation of the temperature is 32^0C (47^0C-15^0C). Absolute measure is possible to compare two sets of data expressed in the same unit i.e. kgs, rupees, etc. In case the two set of data are expressed in different units or in different sizes, the absolute measures of variation are not comparable. In such situations, the measures of relative variation should be used. For example, we would like to compare the variation of temperature about an average which is measured in degrees celsius and the variation of the sale of ‘cold drinks’ given in rupees. This type of distribution variation is obtained as ratio or percentage. We shall now consider in turn each of the four relative measures of variation, which provide a numerical index of the variability of a given data. They are:

i) Range

ii) Quartile Deviation

iii) Mean Deviation, and

iv) Standard Deviation

9.4.1 Range

The Range is the simplest measure of variation. It is defined simply as the difference between the highest value and the lowest value of observation in a set of data. In equation form for absolute measure of range from ungrouped and grouped data, we can say

Range = Highest Value – Lowest Value

In a grouped data, the absolute range is the difference between the upper limit of the highest class and lower limit of the lowest class. The equation form for relative measure of range from ungrouped and grouped data, called coefficient of range is as follows:

$$\text{Coefficient of Range} = \frac{\text{Highest value} - \text{lowest value}}{\text{Highest value} + \text{lowest value}}$$

Let us take an illustration to understand the computation of absolute and relative range.

Illustration-2

The following data relates to the total fares collected on Monday from three different transport agencies.

Transport Agency	Fares on Monday (Rs.)				
A	450	300	300	500	500
B	400	400	400	400	400
C	600	500	200	300	500

Solution: Let us compute the absolute range of the three transport agencies A, B and C.

Range = H.V. – L.V.

A's Range = 500 – 300 = Rs. 200

B's Range = 400 – 400 = Rs. 0

C's Range = 600 – 200 = Rs. 400

The interpretation for the above result is simple. In the above illustration, the variation is nil in case of taxi fare of agency 'B'. While the variation is small in agency 'A' and the variation is high in transport agency C. The coefficient of Range for transport agency 'A' and 'C' is as follows:

$$\text{A's coefficient of R} = \frac{500-300}{500+300} = 0.25 \qquad \text{C's coefficient of R} = \frac{600-200}{600+200} = 0.50$$

Its usefulness as a measure of variation is limited. Since it considers only the highest and lowest values of the data, it is greatly affected by extreme values of the data. Therefore, the range is likely to change drastically from sample to sample. Range cannot be computed in case of open-ended distribution.

In spite of the limitations discussed above, the range is extensively used in specific situations. For instance, it plays an important role in preparing the quality control charts, studying fluctuations in the prices of commodities, price of shares etc. For example maximum gold price and minimum gold price during a specific period. For meteorological department the range is a good indicator for weather forecast to know within what limits the temperature is likely to vary from maximum temperature to minimum temperature in different cities.

Self Assessment Exercise A

The following data relates to the record of time (in minutes) of trucks waiting to unload material.

Company A	0.51	0.68	0.23	0.59	0.93	0.15	0.85
Company B	0.62	0.25	0.36	0.89	1.05	0.20	0.95

Calculate the absolute and relative range and comment on whether you think it is a useful measure of variation.

.....

.....

.....

.....

.....

.....

.....

.....

9.4.2 Quartile Deviation

Quartile Deviation is another measure of variation, also termed as semi-inter quartile range. As we know, quartiles are the factors which divide the distribution into four equal parts i.e., Q_1 (first quartile) gives the value of the 1/4th item and Q_3 (third quartile) gives the value of 3/4th item. The difference between the Q_3 and Q_1 is termed as inter quartile range, when it is divide by two is termed as quartile deviation. It includes the middle 50 per cent of the distribution. As a result, in a given data one quartile at the upper end and another quartile at the lower end are excluded. It is, therefore, unaffected by extreme values. In case of symmetrical distribution Q_1 and Q_3 are equidistant from the median. Where as in asymmetrical distribution Q_1 and Q_3 are not equidistant from median. Symbolically, the absolute measure of quartile deviation may be presented as:

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

The relative measure of Q.D., called coefficient of quartile deviation, is calculated as:

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

It is to be noted that the above formulae (absolute and relative) are applicable to ungrouped data and grouped data as well. Let us take up an illustration to ascertain the value of quartile deviation and coefficient of Q.D.

Illustration-3

The following data relates to the daily expenditure of the students in Delhi University. Calculate quartile deviation and its co-efficient.

Daily expenditure	50-100	100-150	150-200	200-250	250-300	300-350	350-400	400-450	450-500
No. of Students	18	14	21	15	12	13	8	5	2

Solution: For computation of quartile deviation we have to construct the given frequency (No. of students) into cumulative frequency. The procedure is to add the frequency of each class to previous cumulative frequency. In this process, the frequency of the first class is to be taken as the cumulative frequency of that class and the cumulative frequency of the last class is equal to the total frequency (sum of observations) of the given data.

Daily expenditure (x)	No. of Students (f)	Cumulative Frequency
50-100	18	18
100-150	14	32
150-200	21	53
200-250	15	68
250-300	12	80
300-350	13	93
350-400	8	101
400-450	5	106
450-500	2	108

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

Q_1 has N/4th observation i.e., 108/4 = 27th observation which lies in 32 cumulative frequency. So, the value of Q_1 exists in the class 100-150. Now with the help of following formula, the value of Q_1 has to be ascertained.

$$Q_1 = L_1 + \frac{N/4 - c.f}{f} \times i$$

Where, 'L_1' is the lower limit of the Q_1 class, c.f. is the cumulative frequency of the preceding class of Q_1 class 'f' is the frequency of the Q_1, class and 'i' is the class interval. Now we present these values to obtain the result of Q_1.

$$Q_1 = 100 + \frac{27 - 18}{14} \times 50 = \text{Rs.}132.14$$

Q_3, has 3(n/4)th observation i.e., 3(108/4)th = 81th observation. This observation lies in 93 cumulative frequency. So Q_3 lies in the 300-350 class.

$$Q_3 = L_1 + \frac{3N/4 - c.f.}{f} \times i$$

Here, as explained above, L_1; c.f; f; and i are related to Q_3 class

$$\text{Therefore,} \quad Q_3 = 300 + \frac{81-80}{13} \times 50 = \text{Rs.}303.85$$

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{303.85 - 132.14}{2} = \text{Rs.}85.85$$

$$\text{Relative measure of Q.D. i.e., Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{303.85 - 132.14}{303.85 + 132.14} = 0.39$$

From the above data it may be concluded that the variation of daily expenditure among the sample students of DU is Rs. 85.85. The coefficient of Q.D. is 0.39 this relative value of variation may be compared with the other dependent variables of the expenditure like family income of the students, pocket money, habit of spending etc.

Quartile deviation is a useful measure, superior to range, in case of open-ended distribution. It is useful where the distribution is badly skewed, because it is not affected by the extreme values.

Self Assessment Exercise B

The following data shows the profit made by 60 companies in a year.

Compute the quartile deviation and its co-efficient. Do you think this is an appropriate measure for measuring variability? Comment on your opinion.

Profits (in lakhs)	Less than 40	40-45	45-50	50-55	55-60	60-65	65-70	70 & above
No. of Companies	3	12	8	15	10	5	5	2

Solution: Calculation of QD and co-efficient of QD.

Profits (x) (in Rs. lakhs)	No. of Companies (f)	C.f

.....

.....

.....

9.4.3 Mean Deviation

As we know, the important characteristics of an ideal measure of variation is, that it should involve all the data values. Considering this fact, range and quartile deviation are not based on all the observation of the data, as they are positional measures of variation. Consequently, they do not show the scatter around an average, but rather a distance on scale. Where as, the mean deviation overcomes the above weakness by considering all the items of a data set. The mean deviation is the arithmetic mean of absolute difference between the items in a distribution and the average of that distribution. Theoretically, mean deviation can be computed from the mean or the median or the mode. However, in actual practice the mean is frequently used in computing the mean deviation. Under this method, algebraic signs (+, –) are ignored while taking the deviations from average. For un-grouped data, the formula is:

$$\text{M.D. from mean} = \frac{\Sigma|x - \bar{x}|}{N} \quad \text{or M.D. from Median} = \frac{\Sigma|x - Me|}{N}$$

For grouped data the formula is:

$$\text{M.D. from mean} = \frac{\Sigma f|x - \bar{x}|}{N} \quad \text{or M.D. from Median} = \frac{\Sigma f|x - Me|}{N}$$

Where, the two bars indicated that the sign of the difference within the two bars is taken as positive, e.g. $|2-6| = 4$ etc. The co-efficient of Mean deviation for un-grouped and grouped data, the formula is:

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{The average used } (\bar{x} \text{ or Me})}$$

As an illustration, let us consider the following data, which relates to the sales of Company A and Company B during 1995-2001.

Illustration-4

Compute the mean deviation and its co-efficient of the sales of two companies A and B and comment on the result.

Years sales (Rs. in '000)	1995	1996	1997	1998	1999	2000	2001
Company A :	484	572	124	386	920	653	690
Company B :	3554	2645	6524	4255	4940	5450	6890

Solution: For computation of mean deviation, we have to prepare the following table. In this illustration we consider the mean for computation of mean deviation.

Years	Company A		Company B	
	Sales (Rs. in '000) X	Mean = 547 $\lvert x - \bar{x} \rvert$	Sales (Rs. in '000) X	Mean = 4894 $\lvert x - \bar{x} \rvert$
1995	484	63	3554	1340
1996	572	25	2645	2249
1997	124	423	6524	1630
1998	386	161	4255	639
1999	920	373	4940	46
2000	653	106	5450	556
2001	690	143	6890	1996
	3829	1294	34258	8456

Mean Sales of Company 'A' = $\frac{\Sigma X_A}{N_A} = \frac{3829}{7}$ = Rs. 547 thousand

Mean Sales of Company 'B' = $\frac{\Sigma X_B}{N_B} = \frac{34258}{7}$ = Rs. 4894 thousand

Formula for Mean Deviation from Mean = $\frac{\Sigma \lvert x - \bar{x} \rvert}{N}$

M.D. of Company 'A' = $\frac{1294}{7}$ = Rs. 184.9 thousand

M.D. of Company 'B' = $\frac{8456}{7}$ = Rs. 1208 thousand

Co-efficient of M.D.

Company 'A' = $\frac{M.D._A}{Mean_A} = \frac{184.9}{547} = 0.34$

Company 'B' = $\frac{M.D._B}{Mean_B} = \frac{1208}{4894} = 0.25$

The coefficient of mean deviation of company 'A' sales is more than the company 'B' sales. Hence we can conclude there is greater variability in the sales of company 'A'.

The drawbacks of this method are, it may be observed, the algebraic signs (+ or –) of the deviations are ignored. From the mathematical point of view it is unjustifiable as it is not useful for further algebraic treatment. That is the reason mean deviation is not frequently used in business research. The accuracy of the result of mean deviation depends upon the degree of representation of average. Despite of few drawbacks of this measure, it is most useful measure in case of : i) small samples with no-elaborate analysis is required, ii) the reports presented to the general public not familiar with statistical methods, and iii) it has some specific utility in the area of inventory control.

Self Assessment Exercise C

Calculate the mean deviation and its co-efficient from the following data which relates the weekly earnings of the family in an area. What light does it throw on the economical condition of that community and do you justify this measure is a scientific measure of variability? Give your opinion.

Weekly Earnings (Rs.)	0-1000	1000-2000	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
No. of Families	532	704	210	110	32	8	4

Solution: We can also measure the deviations from Median. It is preferred because the average deviation from Median is the least.

Computation of Mean Deviation and its Co-efficient from Median.

Weekly Earnings (Rs.) (x)	Mid-points (x)	No. of families (f)	Less than C.f	$\lvert x - M_e \rvert$	$f\lvert x - M_e \rvert$

.....

.....

9.4.4 Standard Deviation

The Standard deviation is the most familiar, important and widely used measure of variation. It is a significant measure for making comparison of variability between two or more sets of data in terms of their distance from the mean. The mean deviation, in practice, has been replaced by the standard deviation. As discussed earlier, while calculating mean deviation algebraic signs (– / +) are ignored and can be computed from any of the averages. Whereas, in computation of standard deviation signs are taken into account, and the deviation of items, always from mean are taken, squared (instead of ignoring signs) and averaged. So, finally square root of this value is extracted. Thus, standard deviation may be defined as "the square root of the arithmetic mean of the squares of deviations from arithmatic mean of given distribution." This measure is also known as root mean square deviation. If the values in a given data are dispersed more widely from the mean, then the standard deviation becomes greater. It is usually denoted by σ (read as sigma). The square of the standard deviation (σ^2) is called "variance".

As said earlier, it is a type of average deviation of values from mean that is calculated by using the following formulae.

For ungrouped data:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}, \text{ In simple } \sigma = \sqrt{\frac{x^2}{N}}$$

where, x^2 = sum of the squares of deviations $(x - \bar{x})$ and N = No. of observations.

For grouped data:

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})}{N \text{ or } \sum f}}, \text{ In simple } \sigma = \sqrt{\frac{\sum f\, x^2}{N \text{ or } \sum f}}$$

If the collected data are very large, then considering the assumed mean is more convenient to compute standard deviation. In such case, the formula is slightly modified as:

$$\sigma = \sqrt{\frac{\sum f\,(x-\bar{x}_A)^2}{N \text{ or } \Sigma f} - \left[\frac{\sum f\, dx}{N \text{ or } \Sigma f}\right]^2} \times C$$

$$\text{Where, } \bar{x}_A = \text{Assumed mean, } dx = \frac{X - \text{Assumed Mean}}{C}, \quad C = \text{Common factor}.$$

The above formula is applicable only when the class intervals are equal.

Let us take up an illustration to understand the computation of standard deviation from the grouped data, which relate to the profits of 70 companies.

Illustration-5

Profit *(Rs. In Lakh)*	6-10	10-14	14-18	18-22	22-26	26-30
No. of Company	9	11	20	16	9	5

Solution: In order to ascertain the Standard deviation we prepare the following table:

Profit *(Rs. In lakh)*	f	M.V. x	$\frac{x - \bar{x}_A}{c}$ dx	fdx	fdx^2
6-10	9	8	–2	–18	36
10-14	11	12	–1	–11	22
14-18	20	16	0	0	0
18-22	16	20	1	16	16
22-26	9	24	2	18	36
26-30	5	28	3	15	45
	N=70			Σfdx =20	Σfdx^2 = 155

In the above computation, we have taken the mid value "16" as assumed mean (AM), the common factor is 4.

N = 70' Σfdx = 20 Σfdx^2 = 155

$$\sigma = \sqrt{\frac{\sum fdx^2}{N} - \left[\frac{\sum fdx}{N}\right]^2} \times C$$

$$\sigma = \sqrt{\frac{\sum fdx^2}{N} - \left[\frac{\sum fdx}{N}\right]^2} \times C$$

$$= \sqrt{\frac{155}{70} - \left[\frac{20}{70}\right]^2} \times 4 = \sqrt{2.21 - 0.08 \times 4} \quad = \sqrt{2.13 \times 4} \quad = 1.46 \times 4 \quad = 5.84$$

Among all the measures of variation, standard deviation is the only measure possessing the necessary mathematical properties which enhance its utility in advanced statistical work. It is least affected by the fluctuations of sampling. In a normal distribution, $\bar{x} \pm \sigma$ covers 68% of the values whereas $\bar{x}$ ± QD covers 50% values and $\bar{x}$ ± M.D. covers 57% values. This is the reason that standard deviation is called a "Standard Measure".

9.4.5 Coefficient of Variation

The relative measure of standard deviation is the coefficient of variation, denoted by C.V. The absolute measure of standard deviation, discussed above, do not facilitate comparison of two or more data sets in terms of their variability and consistency. However, comparison between such data sets is possible in terms of standard deviation only when the mean and the units of measurement of the data sets are the same. Otherwise, when distributions are measured in the same units but which have different arithematic means and / or distributions are measured in different units then this measure, therefore, is used to compare variability, consistency, and uniformity between two or more sets of data. When C.V. is lesser in the data, it is said to be more consistent or have less variability. On the other hand, the series having higher C.V. has higher degree of variability or lesser consistency. The drawback of this measure is that it is not satisfactorily useful when the mean is close to zero. To provide a standardized measure of variation, we compute the C.V which expresses the standard deviation as a percentage of the mean:

$$C.V. = \frac{\sigma}{X} \times 100$$

Consider the following data which relates to the mean production and standard deviation of Paddy in four states for understanding of the application of C.V.

State	**Mean Production of paddy** ***(In Lakh Tons)***	**Standard Deviation** ***(In lakh tons)***
I	83	9.93
II	40	5.24
III	70	8.12
IV	59	10.89

You may notice that the mean production of Paddy in four states is not equal. In such a situation, to determine which state is more consistent in terms of production, we shall compute the coefficient of variation.

$$C.V. = \frac{\sigma}{X} \times 100$$

$$\text{C.V. of State I} = \frac{9.93}{83} \times 100 = 11.96\%; \quad \text{C.V.of State II} = \frac{5.24}{40} \times 100 = 13.10\%$$

$$\text{C.V. of State III} = \frac{8.12}{70} \times 100 = 11.60\%; \quad \text{C.V. of State IV} = \frac{10.89}{59} \times 100 = 18.46\%$$

It is seen that the standard deviation is low in State II when we compare with the other states. However, since the C.V. is less in State III, it is the more consistent state in production of paddy than the other three states. Among the 4 states, state IV is extremely inconsistent in production of paddy.

Self Assessment Exercise D

A Prospective buyer tested the bursting pressure of a sample of 120 carry bags received from A and B manufactures. The results are tabulated below:

Bursting Pressure *(Kgs)*	**10-12**	**12-14**	**14-16**	**16-18**	**18-20**	**20-22**
No. of bags of A	3	14	30	56	12	5
No. of bags of B	8	16	23	34	24	15

Which manufacturer's bags have the higher average of bursting pressure? Which manufactures would you like to recommend and why? If the buyer does not want to buy bags of more than 16 kg bursting pressure then how does it change your suggestion, if at all?

Solution: Calculation of Standard Deviation and co-efficient of variation.

Busting Pressure (Kgs)	Mid-points (x)	$\frac{x - AM}{i}$ (AM) (dx)	No. of bags (f)	fdx	fdx^2	No. of bags (f)	fdx	fdx^2

.....

.....

.....

.....

.....

.....

9.5 SKEWNESS

The measure of skewness tells us the direction of dispersion about the centre of the distribution. Measures of central tendency indicate only the single representative figure of the distribution while measures of variation, indicate only the spread of the individual values around the means. They do not give any idea of the direction of spread. Two distributions may have the same mean and variation but may differ widely in the shape of their distribution. A distribution is often found skewed on either side of its average, which is termed as asymmetrical distribution. Thus, skewness refers to the lack of symmetry in distribution. Symmetry signifies that the value of variables are equidistant from the average on both sides. In other words, a balanced pattern of a distribution is called symmetrical distribution, where as unbalanced pattern of distribution is called asymmetrical distribution.

A simple method of finding the direction of skewness is to consider the tails of a frequency polygon. The concept of skewness will be clear from the following three figures showing symmetrical, positively skewed and negatively skewed distributions.

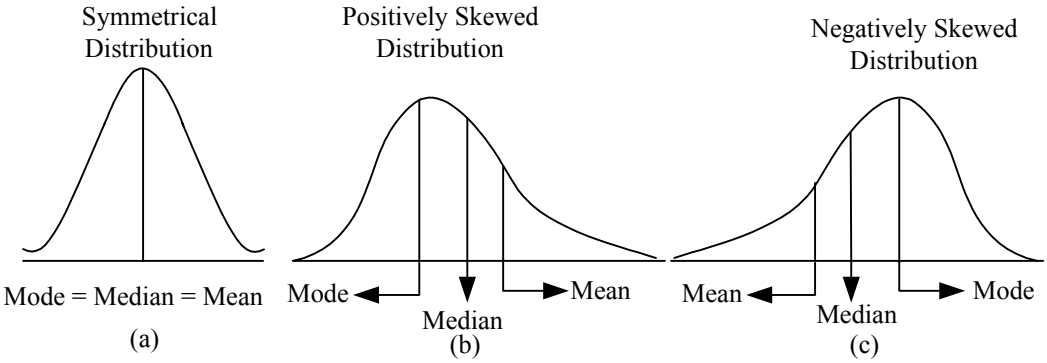

Fig.9.1

Carefully observe the figures presented above and try to understand the following rules governing them.

It is clear from Figure 9.1 (a) that the data are symmetrical when the spread of the frequencies is the same on both sides of the middle point of the frequency polygon. In this case the value of mean, median, and mode coincide i.e., Mean = Median = Mode.

When the distribution is not symmetrical, it is said to be a skewed distribution. Such a distribution could be either positively skewed or negatively skewed. In Figure (b), when there is a longer tail towards the right hand side of the centre of distribution, the skewness is said to be “Positively Skewed”. In such a situation, Mean > Median > Mode.

In Figure (c), when there is a longer tail towards the left hand side of the centre, then the skewness is said to be ‘Negatively Skewed’. In such a case, Mean < Median < Mode.

It is seen that, in positively skewed distribution, dispersal of individual observations is greater towards the right of the central value. Where as in a negatively skewed distribution, a greater dispersal of individual observations is towards the left of the central value. We can say, therefore, the concept of Skewness not only refers to lack of symmetry in a distribution but also indicates

the magnitude as well as the direction of skewness in a distribution. The relationship of mean, median and mode in measuring the degree of skewness is that, for a moderately symmetrical distribution the interval between the mean and the median is approximately $1/3^{rd}$ of the interval between the mean and mode.

Tests of Skewness

In the light of the above discussion, we can summerise the following facts regarding presence of skewness in a given distribution.

1) The mean, median, and mode are not identical.
2) The total of deviations are not zero from median or mode i.e. $\Sigma(X - Me)$ or $\Sigma(X - Mo) \neq 0$.
3) Frequencies on both sides of the mode are not equal
4) The distance from the Median to the quantities are not equal i.e., $(Q_3 - Me)$ is not equal to $(Me - Q_1)$.
5) The curve of distribution is not bell shaped. This means the two halves of the curve from Median or Mode do not coincide in a perfect manner.

9.6 RELATIVE SKEWNESS

The relative measure of skewness is termed as Coefficient of Skewness, It is useful in making a comparison between the skewness in two or more sets of data. There are two important methods for measuring the coefficient of skewness. They are: 1) Karl Pearson's coefficient of skewness. 2) Bowley's coefficient of skewness.

Let us discuss these two methods. Study carefully to understand the computation of co-efficient of skewness.

i) Karl Pearson's Coefficient of Skewness: (denoted as SKp.)

This co-efficient of skewness, is obtained by dividing the difference between the mean and the mode by the standard deviation. Thus the formula of Pearson's coefficient of skewness is:

$$SKp = \frac{\overline{X} - M_O}{\sigma}$$

This method computes the co-efficient skewness by considering all the items of the data set. The value of variation usually varies in value between the limits ± 3.

If mode is ill-defined and cannot be easily located then using the approximate empirical relationship between mean, median, and mode as stated in Unit-8 section 8.3.5, (mode = 3 median – 2 mean) the coefficient of skewness can be determined by the removal of the mode and substituting median in its place. Thus the changed formula is:

$$Sk_p = \frac{3\,(\text{Mean} - \text{Median})}{\sigma}$$

Let us consider the following data to understand the application of Karl Pearson's formula for measuring the co-efficient of skewness.

Illustration-6

The following measures are obtained from the profits of 100 shops in two different regions. Calculate Karl Pearson's co-efficient of skewness and comment on the results.

Region I : $\overline{X} = 16.62;\ Mo = 18.47;\ \text{and}\ \sigma = 3.04$

Region II : $\overline{X} = 45.56;\ \ Mo = 36.94;\ \ \text{and}\ \sigma = 17.71$

Note that, we have already learnt the computation of $\overline{x}$, M_0 in Unit 8 and σ in this Unit.

Solution: $\text{Karl Pearson's formula} : (SK_p) = \dfrac{\overline{X} - M_0}{\sigma}$

$$\text{Coefficient of skewness for Region I} = \frac{16.62 - 18.47}{3.04} = -0.61$$

$$\text{Coefficient of skewness for Region II} = \frac{45.36 - 36.94}{17.71} = 0.49$$

Based on the results we can comment on the distributions of the two regions as follows: The coefficient of skewness for Region-I is negative, while that of Region - II is positive. Hence the earnings of profit in Region I is more skewed. Since the result in Region-I, indicates that the distribution is negatively skewed, there is a greater concentration towards higher profits. In case of Region-II the value of coefficient of skewness indicates that the distribution is positively skewed. Therefore there is a greater concentration towards lower profits.

Let us consider another illustration to understand the application of Pearson's alternative formula for co-efficient of skewness, when it is not possible for the mode to be located in a distribution.

Illustration-7

The following statistical measures are given from a data of a factory before and after the settlement of wage dispute. Calculate the Pearson's co-efficient of skewness and comment.

Particulars	Before Settlement	After Settlement
No. of workers	1200	1175
Standard deviation (Rs.)	5.9	4.95
Mean wage (Rs.)	22.8	24.0
Median wage (Rs.)	24.0	23.0

Solution: It is understood that the mode is ill-defined in the given data. Hence, to compute the Pearson's coefficient of skewness, the following alternative formula is used here.

$$\text{Kal Pearson's Co-efficient of Skewness } (SK_p) = \frac{3\,(\text{Mean} - \text{Median})}{\sigma}$$

a) Before settlement of wage dispute: $SK_p = \frac{3(22.8-24.0)}{5.9} = \frac{-3.6}{5.9} = -0.61$

b) After settlement of wage dispute: $SK_p = \frac{3(24-23)}{4.9} = \frac{3}{4.95} = 0.61$

From the above calculated values of coefficient of skewness, under different situations, we may comment upon the nature of distribution as follows:

Before the settlement of dispute the distribution was negatively skewed and hence there is a greater concentration of wages towards the higher wages. Whereas it was positively skewed after the settlement of dispute, which reveals that even though the mean wage of workers has increased after the settlement of disputes (before settlement wages were 1,200 × 22.8 = Rs. 27360. After settlement total wages were 1175 × 24 = Rs. 28,200). The workers who were getting low wages are getting considerably increased wages after settlement of their dispute, while wages of the workers getting high wages before settlement had fallen.

We can also comment on the level of uniformity in the distribution of wages by computing co-efficient of variation which we have studied in this unit, Section 9.4.5. Hence, let us compute the variation to study the state of uniformity in wages in both the circumstances.

$$C.V. = \frac{\sigma}{X} \times 100$$

a) Before settlement the coefficient of variation $= \frac{5.9}{22.8} \times 100 = 25.88\%$

b) After settlement the coefficient of variation $= \frac{4.95}{24.0} \times 100 = 20.62\%$

Based on the computed values of variation, it may be concluded that there is sufficient evidence that there is lesser inequality in the distribution of wages after settlement of the dispute. It means that there was a greater scattered in wage payment before the dispute was settled.

Self Assessment Exercise E

A survey was conducted on random basis by a Television manufacturing company to enquire the maximum price at which persons would be willing to purchase the colour T.V. The following table gives the stated price (in thousand Rs.) by 150 respondents.

Price of T.V. *(Thousand Rs.)*	8-10	10-12	12-14	14-16	16-18	18-20
No. of persons	19	23	28	40	26	14

Calculate Karl Pearson's co-efficient of skewness and interpret the result.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Bowley's Measure of Co-efficient of Skewness

Bowley's method for coefficient of skewness (SK_B) is derived from quartile values and for this reason it is useful in case of open-ended distribution, where extreme values are presented and/or class intervals are unequal in the collected data or the median and quartile values only are available. In such situations, formula for coefficient of skewness developed by Prof. Bowley is more appropriate. It is expressed as:

$$SK_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}, \quad \text{Alternatively;}$$

$$SK_B = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

It is to be noted that, in an asymmetrical distribution the value of this co-efficient of skewness lies between ± 1. The criticism against this measure is that it does not take all the items of the data into account. It is based on control of 50% of the data and it ignores 25% of the data below Q_1 and 25% of the data above Q_3. Thus, this method is also termed as Quartile Co-efficient of skewness. Since, this method is based only on the middle 50% of the distribution, there is a possibility that SK_B may be negative even while SKp is positive. However this is a useful measure when variability of the distribution is computed by using the method of quartile deviation.

Let us consider the following illustration to understand the concept of Bowley's co-efficient of Skewness.

Illustration-8

The following values were computed in an open-ended distribution relating to sales of a product. Compute the co-efficient of skewness.

$Q_1 = 62 \qquad Q_2 = 141 \qquad Q_3 = 190$

Solution: When quartiles are given, as we discussed earlier (Section 9.4.2), Bowley's conept is appropriate to obtain the value of relative skewness. Here, the median (Q_2 value is exactly equal to the value of Median) can also be denoted as Q_2.

Bowley's coefficient of skewness (SK_B) = $\dfrac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$

Bowley's co-efficient of skwness $SK_B = \dfrac{190 + 62 - (2 \times 141)}{190 - 62} = \dfrac{252 - 282}{128} = -0.23$

since the distribution is slightly negatively skewed there is greater concentration of the sales towards higher sales than the lower sales of the distribution.

Self Assessment Exercise F

From the following information regarding the payment of Commission to salesmen in two companies (X Ltd. And Y Pvt. Ltd)

Calculate Bowley's co-efficient of skewness and find out which company is more homogeneous in payment of commission and which is more a skew. How do you justify that Bowley's measure is appropriate?

X Ltd		Y Pvt. Ltd.	
Payment of Commission in Rs.	No. of Salesmen	Payment of Commission in Rs.	No. of Salesmen
200-250	50	400-450	20
250-300	85	450-500	42
300-350	67	500-550	50
350-400	58	550-600	22
400-450	16	600-650	16
450-500	7	650-700	9

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

9.7 LET US SUM UP

In this Unit, we have studied how the concepts of variation and Skewness are useful for describing the data more meaningfully. Variation is a measure of scatter or spread of items around the central values. Variation is calculated to examine the extent to which the items vary from some central values. Thus the average is more meaningful, if it is examined in the light of variation. Relative measures are obtained as ratios and percentages and are used to compare variability in two or more sets of data. The mean and the standard deviation may be the same in two different distributions, but it does not imply that the distributions are the same. Hence, there is another measure called the measure of variation. They are range, quartile deviation, mean deviation and standard deviation. We also discussed the concept of coefficient of variation, which is used to compare relative variation of different sets of data.

Through skewness, we study the shape of the distribution, i.e., whether the distribution is symmetrical or asymmetrical. Symmetrical distribution means the frequency distribution that forms a balanced pattern on both sides of the mean, median, and mode.

In such a distribution mean, median, and mode are equal and they lie at the centre of the distribution. In contrast, asymmetrical distribution means unbalanced pattern of frequency distribution, called as 'skewed' distribution. Skewed distribution may be positively skewed or negatively skewed. In a positively skewed distribution the mean is greater than mode and median ($\bar{x}$ > Me > Mo) and has a long tail on the right hand side of the data curve. On the other hand, in a negatively skewed distribution the mode is greater than the mean and median (Mo > M_e > $\bar{x}$) and has a long tail on the left hand side of the data curve. In a skewed distribution, the relationship of Mean and median is that the interval between both is approximately 1/3rd of the interval between the mean and mode. Based on this relationship the degree of skewness is measured. There are two formulae we use for measuring the coefficient of skewness, which are called relative measures of skewness, proposed by Karl Pearson and Bowley. Bowley's formula is normally applied when the data is an open-end type or/and the classes are unequal.

9.8 KEY WORDS

Asymmetry : A characteristic of a distribution in which the values of variables are not equidistant from the average on both sides.

Co-efficient of Skewness : It makes comparison between the skewness in two or more data sets.

Co-efficient of Variation : A relative measure of variation, comparable across distributions, that is, the ratio of the standard deviation to mean expressed as a percentage.

Mean Deviation : The arithmetic mean of the absolute values of the deviations from some measure of central tendency (Mean, Median or Mode).

Measure of Variation : A measure describing how the observations in a distribution are spread or scattered.

Quartile Deviation : It is one half of the difference between the upper quartile (Q_3) and the Lower quartile.(Q_1).

Range : It is the difference between the highest value and the lowest value of observations.

Standard Deviation : The square root of the Arithmetic mean of squares of deviations from Arithmetic mean of the data set.

Skewness : It refers to the lack of symmetry in distribution.

Symmetry : A characteristic of a distribution in which the values of variables are equidistant from the average on both sides.

Variation : The degree to which numerical data tend to spread about an average value.

Variance : The square of standard deviation.

9.9 ANSWERS TO SELF ASSESSMENT EXERCISES

A) Range : Company A = 0.78 Minutes; Company B = 0.85 Minutes,

Co-efficient of Range: Company A = 0.72; Company B = 0.68. It is not particularly useful, because in case of A Company, all the rest of the items, except two fall between 0.51 minutes and 0.93 minutes. In case of Company B also the items except three, fall between 0.62 minutes and 1.05 minutes. The range greatly overstates the typical variability, because it is determined by two extreme values in the data set.

B) $Q_1 = 45$, $Q_3 = 57.92$; Q.D = 6.46; co-efficient of Q.D = 0.063.

Yes. Even though, the Quartile deviation is regarded as a measure of partition and may not satisfy the test of a good measures of variation, it is an approximate method in specific situations where the data is in the form of open-ended classes. It is also useful where the distribution is badly skewed, because it is unaffected by the extreme values.

C) Median = Rs. 1380.68

Mean deviation = Rs. 722.

Co-efficient of Mean Deviation = 0.52.

The Median earnings of the 1600 families is Rs. 1381. It reveals that 50% of the families are earning between Rs. 1,000 to Rs. 2,000. It is to be noted that very few (44 families out of 1,600 families) fall in the last three classes of higher-earning groups.

This is not a scientific measure of variability because while taking the deviations algebric signs are ignored Therefore, it is not capable of further algebraic treatment.

Infact, this measure of variability gives us best results when deviations are taken from Median, but Median is not a satisfactory measure when the dispersion in a distribution is very high. It is also not appropriate for large samples.

D) Manufacturer A: $\bar{x}$ = 16.25 kgs; σ = 2.07 kgs; CV = 12.74%

Manufacturer B: $\bar{x}$ = 16.58 kgs; σ = 2.81 kgs; CV = 16.95%

Since the mean bursting pressure of manufacturer B's bags is higher, these bags may be regarded more standard. However, the bags of manufacturer A may be suggested for purchase as these bags of manufacturer A are more consistant because CV is significantly lesser than the bags of manufacturer B.

If the buyer would not like to buy bags having more than 16 kgs. bursting pressure then:

$\bar{x}_A = 14.15$; $\sigma_A = 0.74$; $CV_A = 5.23\%$

$\bar{x}_B = 13.64$; $\sigma_B = 1.12$; $CV_B = 8.21\%$

In case the buyer would not like to buy bags having more than 16 kgs bursting pressure, then the average bursting pressure of manufacturer A's

bags is higher than manufacturer B. The co-efficient of variation is also much lesser in case of manufacturer A than manufacturer B. Hence, in this case, we may suggest to buy from manufacturer A.

E) $\bar{x}$ = 13.97; M_0 = 14.92; σ = 2.9, SK_p = – 0.32. Since SK_p is – 0.32, the distribution is asymmetrical and negatively skewed. It is true because the mode is greater than the mean. The absolute measure of skewness i.e. $\bar{x} - M_0$ is – 0.95 (13.97–14.92). Such an asymmetrical distribution graphically would tend to tail off towards the left side.

F) Company X : Q_1 = 262.21; M_e = 304.85; Q_3 = 358.84; SK_B = 0.12.

Company Y : Q_1 = 473.51; M_e = 517.5; Q_3 = 556.48; SK_B = – 0.06.

9.10 TERMINAL QUSTIONS AND EXERCISES

1) What do you understand by "Variation"? Discuss the significance of measuring variability for data analysis.
2) When would you use the range and standard deviation as a measure of variation? Explain with suitable illustrations.
3) Explain in what ways measures of variation supplement measures of central tendency.
4) Explain the concept of skewness. How does it help in analyzing the data?
5) Distinguish between variation and skewness. What are the objectives of measuring them?
6) The following table is related to the daily temperatures recorded in a city in a year. Calculate Range and Quartile deviation and which measure of variation do you suggest to take a decision on the type of garments to be produced by a garment factory. Justify your suggestion.

Temperature °C	–30 to –20;	–20 to –10;	–10 to 0	0 to 10;	10 to 20
No. of days	38	190	65	42	30

7) Calculate mean deviation from the following distribution.

Profits *(in lakhs)*	No. of firms	Profits *(in lakhs)*	No. of firms
15-25	14	45-55	14
25-35	28	55-65	32
35-45	56	65-70	26

8 A transport agency had tested the tyres of two brands A and B. The results are given in the following table below.

Life *(thousand units)*	Brand A	Brand B
15-20	6	8
20-25	15	8
25-30	10	22
30-35	16	17
35-40	13	12
40-45	9	6
45-50	11	0

i) Which brand of tyres do you suggest to the transport agency to use on their fleet of trucks?

8) In a manufacturing firm, four employees on the same job show the following results over a period of time.

	A	B	C	D
Mean time of completing the Job (minutes)	61	70	83	80.5
Variance (σ^2)	64	81	121	100

i) Which employee appears to be more consistent in the time he/she requires to complete the job?
ii) Which employee appears to be faster in completing the job?
iii) Which measure did you select to answer part (i) and why?

9) The following table relates to the marks obtained at Engineering exams and CAT examination.

i) Find which group is more homogeneous in intelligence?
ii) Which group is more skewed and why?

Engineering examination		CAT examination	
Marks	No. of students	Marks	No. of students
50-100	15	750-800	45
100-150	40	800-850	80
150-200	45	850-900	78
200-250	20	900-950	55
250-300	14	950-1000	12

10) The following Table gives the No. of defects per product and its frequency.

No. of defects per product	Frequency
Under 15	32
15-20	50
20-25	75
25-30	130
30-35	145
35-40	105
40-45	85
45-50	50
50 and above	20

i) What are the problems you may face in computing standard deviation from the above data?
ii) Compute Bowley's co-efficient of skewness and comment on its value.
iii) Do you agree that the suggested method for measuring skewness is an appropriate method? Give reasons of your opinion.

11) The following information was obtained from records of a factory relating to the wages, before and after settlement of wages.

Particulars	Before settlement of dispute	After settlement of dispute
No. of workers	515	507
Mean wage (Rs.)	49.40	51.73
Meadian Wage (Rs.)	52.5	50.00
Standard Deviation of wages	10.00	11.00

i) Give as much information as you can about the distribution of wages.
ii) Comment on the gain and loss from the point of view of workers and that of the factory's management.

12) Students' ages in the regular (conventional) M.Com. Programme and the part-time (distance) programme of a University and an Open University are given by the following two samples:

Regular M.Com	20	24	18	22	26	25	21	28	23	29
Distance M.Com	24	29	40	46	34	27	31	28	38	23

If homogeneity of the class is a positive factor in learning, use a measure of relative variation to suggest which of these two groups will be easier to teach.

Note: These questions/exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the university for assessment. These are for your practice only.

9.11 FURTHER READING

The following text books may be used for more indepth study on the topics dealt with in this unit.

Clark, T.C. and E.W. Jordon, 1998, *Introduction to Business and Economic Statistics*, South-Western Publishing Co.

Gupta, S.P. and Gupta, M.P. 2000, *Business Statistics*, Sultan Chand & Sons : New Delhi.

Hooda, R.P. 2001, *Statistics for Business and Economics*, Macmillian India Ltd. New Delhi.

Richard I. Levin and David S. Rubin, 2000, *Statistics for Management*, Prenctice-Hall of India Pvt. Ltd. New Delhi.