Chapter 4

> *There never was in the world two opinions alike, no more than two hairs or two grains; the most universal quality is diversity.*
>
> —Michel de Montaigne

> *I feel like a fugitive from the law of averages.*
>
> —Bill Mauldin

Measures of Dispersion

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- provide the importance of the concept of variability (dispersion).
- measure the spread or dispersion, understand it, and identify its causes to provide a basis for action.

4.1 INTRODUCTION

The measures of central tendency describe that the values in the data set tend to spread (cluster) around a central value called *average*. But these measures do not reveal how these values are spread (dispersed or scattered) on each side of the central value. Just as central tendency can be measured by a number in the form of an average, the amount of variation (dispersion, spread or scatter) among the values in the data set can also be measured. The dispersion of values is indicated by the extent to which these values tend to spread over an interval rather than cluster closely around an average.

The statistical techniques to measure the extent to which values in the data set tend to spread are of two types:

(i) Techniques that are used to measure the extent of variation or deviation of each value in the data set from a measure of central tendency, usually the mean or median. Such statistical techniques are called *measures of dispersion* (or *variation*).

(ii) Techniques that are used to measure the direction (away from uniformity or symmetry) of variation in the distribution of values in the data set. Such statistical techniques are called *measures of skewness* (see Chapter 5).

Identifying the causes and then measuring the dispersion is useful to draw statistical inference (estimation of parameter, hypothesis testing, forecasting and so on). A small dispersion among values in the data set indicates that values in the data set are clustered closely around the mean, implying that the mean is a reliable average. Conversely, if values in the data set are widely clustered around the mean, then this implies that the mean is not a reliable average, i.e. mean is not representative of the data.

The symmetrical distribution of values in two or more data sets may have same variation but differ in terms of A.M. as shown in Fig 4.1. On the other hand, two or more data sets may have the same A.M. values but differ in variation as shown in Fig. 4.2.

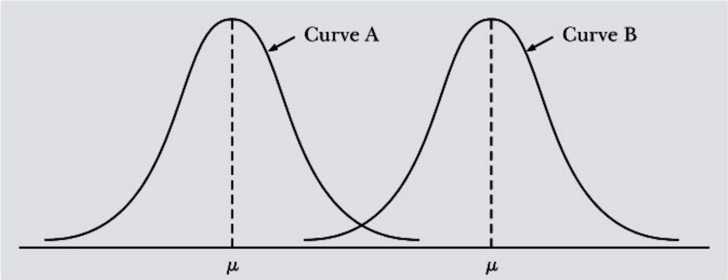


Figure 4.1
Symmetrical Distributions with Unequal Mean and Equal Standard Deviation

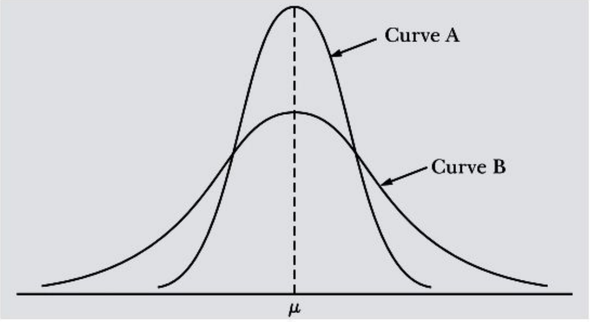


Figure 4.2
Symmetrical Distributions with Equal Mean and Unequal Standard Deviation

Illustration Suppose over the six-year period the net profits (in percentage) of two firms are as follows:

Firm 1	:	5.2,	4.5,	3.9,	4.7,	5.1,	5.4
Firm 2	:	7.8,	7.1,	5.3,	14.3,	11.0,	16.1

Since average amount of profit is 4.8 per cent for both firms, therefore both the firms are equally good and that a choice for investment purposes must depend on other considerations. However, in case of Firm 2, net profit values are varying from 5.3 to 16.1 per cent, i.e., difference among the values is more while the net profit values of Firm 1 are varying from 3.9 to 5.4 per cent, i.e., difference among the values is less as compared to Firm 2. In other words, net profit values in data set 2 are spread more than those in data set 1. This implies that the performance of Firm 1 is consistent as compared to Firm 2. Consequently, for investment, a comparison of the average (mean) profit values alone should not be sufficient.

4.2 SIGNIFICANCE OF MEASURING DISPERSION

The following are some of the purposes for which measures of variation are needed.

- ***Test the reliability of an average:*** Measures of variation help to understand the extent an average represents the characteristic of a data set. If the extent of dispersion of values is less on each side of an average value, then it indicates high uniformity among values in the distribution. On the other hand, if the variation is large, then it indicates a lower degree of uniformity among values in the data set, and the average value may be unreliable.
- ***Control the variability:*** Measuring variation helps to identify the nature and causes of variation. Such information is useful in controlling the variations. According to Spurr and Bonini, *In matters of health, variations in, body temperature, pulse beat and blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programmes.* In social science, the measurement of 'inequality' of distribution of income and wealth requires the measurement of variability.

- ***Compare two or more sets of data with respect to their variability:*** Measures of variation help in comparing variation in two or more sets of data with respect to their uniformity or consistency. For example, (i) measurement of variation in share prices and their comparison with respect to different companies over a period of time, and (ii) measurement of variation in the length of stay of patients in a hospital helps to set staffing levels, number of beds, number of doctors, and other trained staff, patient admission rates and so on.
- ***Facilitate the use of other statistical techniques:*** Measures of variation facilitate the use of other statistical techniques such as correlation and regression analysis, hypothesis testing, forecasting, quality control and so on.

4.2.1 Requisites for a Good Measure of Variation

Certain essential requisites that help in identifying the merits and demerits of individual measure of variation are as follows:

(i) Should be based on all the values (elements) in the data set.
(ii) Should be calculated easily, quickly and accurately.
(iii) Should be unaffected by the fluctuations in sample size and also by outliers.
(iv) Should be further mathematical or algebraic changes are possible.

4.3 CLASSIFICATION OF MEASURES OF DISPERSION

Measures of dispersion (or variation) based on the purpose of measuring are classified into two categories:

1. **Absolute measures:** These measures are described by a number (or value) to represent the amount of variation (or difference) among values in a data set. Such a value is expressed in the same unit of measurement such as rupee, inch, foot, kilogram, ton, etc., as the values in the data set. Such measures help in comparing two or more sets of data in terms of absolute magnitude of variation, provided variable values are expressed in the same unit of measurement and have almost the same average value.
2. **Relative measures:** These measures are described as the ratio of a measure of absolute variation to an average and is termed as *coefficient of variation*. The word 'coefficient' means a number that is independent of any unit of measurement. While computing the relative variation, the average value used as base should be the same from which the absolute deviations were calculated.

Another classification of the measures of variation is based on the method used for their calculations:

(i) Distance measures
(ii) Average deviation measures

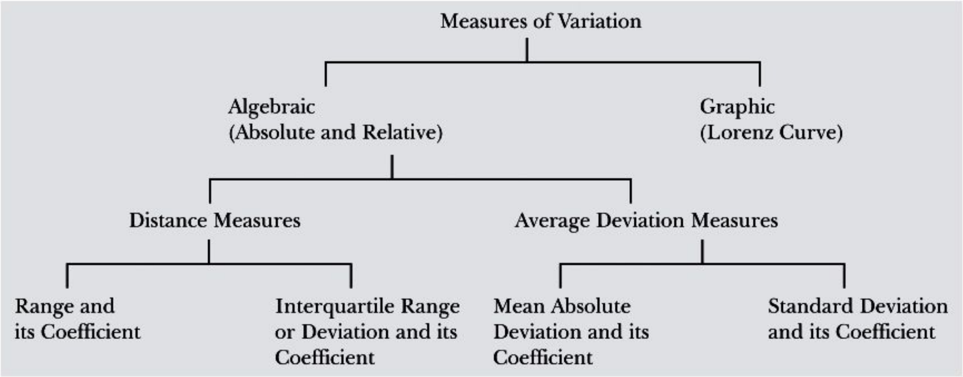

Measures of Variation
Algebraic (Absolute and Relative)
Graphic (Lorenz Curve)
Distance Measures
Average Deviation Measures
Range and its Coefficient
Interquartile Range or Deviation and its Coefficient
Mean Absolute Deviation and its Coefficient
Standard Deviation and its Coefficient

Figure 4.3 Classification of Measures of Variation

The **distance measures** describe the spread or dispersion of values of a variable in terms of difference among values in the data set. The **average deviation measures** describe the average deviation for a given measure of central tendency.

The classification of various measures of dispersion (variation) is shown in Fig. 4.3.

4.4 DISTANCE MEASURES

The distance measures are further classified into two following categories:

(i) Range
(ii) Interquartile deviation

4.4.1 Range

The calculation of range as a measure of dispersion is based on the location of the largest and the smallest values in the data. Thus, **range** is defined as the difference between the largest and lowest observed values in a data set. In other words, it is the length of an interval which covers the highest and lowest observed values in a data set and measures the dispersion or spread within the interval.

Range: A measure of variability, defined to be the difference between the largest and lowest values in the data set.

$$\begin{aligned}\text{Range (R)} &= \text{Highest value of an observation} - \text{Lowest value of an observation}\\ &= \text{H} - \text{L} \qquad (4\text{-}1)\end{aligned}$$

For example, if the smallest value of an observation in the data set is 160 and largest value is 250, then the range is 250 – 160 = 90.

For grouped frequency distribution of values in the data set, the range is the difference between the upper class limit of the last class and the lower class limit of first class. In this case, the range obtained may be higher than as compared to ungrouped data because class limits are extended slightly beyond the extreme values in the data set.

Coefficient of Range

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

$$\text{Coefficient of range} = \frac{\text{H} - \text{L}}{\text{H} + \text{L}} \qquad (4\text{-}2)$$

Example 4.1: The following are the sales figures of a firm for the last 12 months

Months	:	1	2	3	4	5	6	7	8	9	10	11	12
Sales (₹ '000)	:	80	82	82	84	84	86	86	88	88	90	90	92

Calculate the range and coefficient of range of sales for the last 12 months.

Solution: Given that, H = 92 and L = 80. Therefore,

$$\text{Range} = \text{H} - \text{L} = 92 - 80 = ₹12$$

and

$$\text{Coefficient of range} = \frac{\text{H} - \text{L}}{\text{H} + \text{L}} = \frac{92 - 80}{92 + 80} = \frac{12}{172} = 0.069$$

Example 4.2: The following data show the waiting time (to the nearest 100th of a minute) of telephone calls to be matured:

Waiting Time	*Frequency (Minutes)*	*Waiting Time*	*Frequency (Minutes)*
0.10 – 0.35	6	0.88 – 1.13	8
0.36 – 0.61	10	1.14 – 1.39	4
0.62 – 0.87	8		

Calculate the range and coefficient of range for telephone calls to be matured.

Solution: Given that, H = 1.39 and L = 0.10. Therefore,

$$\text{Range} = \text{H} - \text{L} = 1.39 - 0.10 = 1.29 \text{ min}$$

and
$$\text{Coefficient of Range} = \frac{\text{H} - \text{L}}{\text{H} + \text{L}} = \frac{1.39 - 0.10}{1.39 + 0.10} = \frac{1.29}{1.49} = 0.865$$

Advantages, Disadvantages and Applications of Range

Advantages

1. The measurement of range is independent of the measure of central tendency and easy to calculate and understand.
2. The knowledge of range is useful in cases where the purpose is only to know the extent of extreme variation, such as quality control limits, temperature, rainfall and so on.

Disadvantages

1. The calculation of range is based on only two values—largest and smallest in the data set. Thus, the value of range is influenced by two extreme values and completely independent of the other values. For example, range of two data sets {1, 2, 3, 7, 12} and {1, 1, 1, 12, 12} is 11 but the two data sets differ in terms of overall dispersion of values
2. The value of range is sensitive to changes in sample size, i.e., different samples of the same size from the same population may have different ranges.
3. Range cannot be computed for open-ended frequency distributions because no highest or lowest value exists in such cases.
4. The value of range does not describe variation among values between highest and lowest value in a given data set. For example, each of the following data set

Set 1	:	9	21	21	21	21	21	21	21
Set 2	:	9	9	9	9	21	21	21	21
Set 3	:	9	10	12	14	15	19	20	21

has a range of 21 – 9 = 12, but the variation of values between the highest and lowest values is different in each case.

Applications

1. The knowledge of range is useful in the study of small variations among values in a data set. Variation (fluctuation) in share prices and other commodities that are very sensitive to price changes from one period to another may easily be understood by calculating the range of such variations (fluctuations).
2. Quality control is exercised by preparing suitable *control charts*. The control charts are prepared on setting an upper control limit (range) and a lower control limit (range) within which quality of products is acceptable. The variation in the quality beyond these *ranges* requires necessary remedial actions.
3. For weather forecasts, the knowledge of range (difference between maximum and minimum temperature or rainfall) is important.

4.4.2 Interquartile Range or Deviation

Interquartile Range: A measure of variability, defined to be the difference between the quartiles Q_3 and Q_1.

The limitations or disadvantages of range can partially be overcome by using another measure of variation called **Interquartile Range or Deviation (IQR)**. The IQR measures the spread within middle half of the values in the data set so as to minimize the influence of outliers (extreme values) in the calculation of range. Since a large number of values in the data set lie in the central part of the frequency distribution, it is necessary to study the **Interquartile Range** (also called mid-spread).

To compute IQR, data set is divided into four parts each of which contains 25 per cent of the observed values. Thus, *interquartile range* is *a measure of dispersion or spread of values in the data set between the third quartile, Q_3 and the first quartile*, Q_1. In other words, the *interquartile range or deviation* is the range for the middle 50 per cent of the data set. The concept of IQR is shown in Fig. 4.4:

$$\text{Interquartile range (IQR)} = Q_3 - Q_1 \qquad (4\text{-}3)$$

Half the distance between Q_1 and Q_3 is called the *semi-interquartile range* or the *quartile deviation* (QD).

$$\text{Quartile deviation (QD)} = \frac{Q_3 - Q_1}{2} \qquad (4\text{-}4)$$

The median is not necessarily midway between Q_1 and Q_3, although it is true for a symmetrical distribution. The median and quartiles divide the data set into equal numbers of values but do not necessarily divide the data into equally wide intervals.

The ***quartile deviation (QD)*** measures the average range of 25 per cent of the values in the data set. It is computed by taking an average of the middle 50 per cent of the observed values rather than 25 per cent part of the values in the data set.

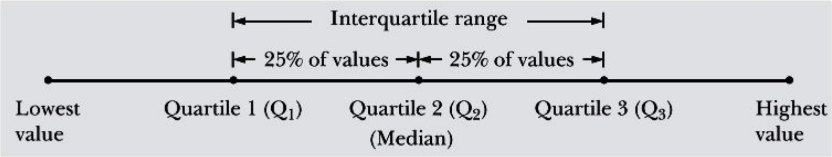


Figure 4.4
Interquartile Range

In a symmetrical distribution, the two quartiles Q_1 and Q_3 are at equal distance from the median, i.e., Median – Q_1 = Q_3 – Median. Thus, *Median ± Quartile Deviation* covers exactly 50 per cent of the observed values in the data set.

A smaller value of quartile deviation indicates high uniformity (or less variation) among the middle 50 per cent values around the median value. On the other hand, a high value of quartile deviation indicates large variation among the middle 50 per cent values.

The median and quartiles divide the data set into equal parts of values but not necessarily into equally wide intervals.

Coefficient of Quartile Deviation

Since quartile deviation is an absolute measure of variation, therefore its value gets affected by the size and number of observed values in the data set. Thus, Q.D. of two or more than two data sets may differ. Due to this reason, to compare the degree of variation in different data sets, we compute the relative measure corresponding to Q.D., called the **coefficient of Q.D.** as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad (4\text{-}5)$$

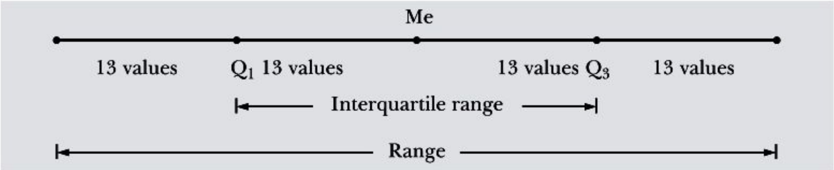
Example 4.3: Following are the responses from 55 students to the question about how much money they spent every day.

55	60	80	80	80	85	85	85	90	90	90
90	92	94	95	95	95	95	100	100	100	100
100	100	105	105	105	105	109	110	110	110	110
112	115	115	115	115	115	120	120	120	120	120
124	125	125	125	130	130	140	140	140	145	150

Calculate the range and interquartile range and interpret your result.

Solution: Since number of responses are 55 — an odd number, therefore median of the given values in the data set is: (55+1)/2 = 28th value which is 105. This means there are 27 values at or below 105 and another 27 at or above 105.

The lower quartile Q_1 = (27+1)/2 = 14th value from bottom of the data, i.e. Q_1 = 94 and upper quartile is the 14th value from the top, i.e. Q_3 = 120. The 55 values have been partitioned as follows:



The interquartile range (IQR) is: 120 – 94 = ₹26, while the range is, R = 150 – 55 = ₹95. The middle 50 per cent values of data set fall in a narrow range of only ₹26. This means responses are densely clustered near the centre of the data and more spread towards the extremes. For instance, lowest 25 per cent of the students had responses in the interval 55 to 94, i.e. ₹39, while the next 25 per cent had responses in the interval 94 to 105, i.e. ₹11. Similarly, the third quarter had responses in the interval 105 to 110, i.e. ₹5, while the top 25 per cent had responses in the interval 120 to 150, i.e. ₹30.

Example 4.4: Use an appropriate measure to evaluate the variation in the following data:

Farm Size (acre)	*No. of Farms*	*Farm Size (acre)*	*No. of Farms*
below 40	394	161–200	169
41–80	461	201–240	113
81–120	391	241 and above	148
121–160	334		

Solution: Since first and last intervals in the frequency distribution are open-end class intervals, Q.D. is an appropriate measure to evaluate variation. The computation of Q.D. is shown in Table 4.1.

Table 4.1 Calculations of Quartile Deviation

Farm Size (acre)	*No. of Farms*	*Cumulative Frequency (cf) (less than)*
below 40	394	394
41–80	461	855 ← Q_1 class
81–120	391	1246
121–160	334	1580 ← Q_3 class
161–200	169	1749
201–240	113	1862
241 and above	148	2010
	2010	

Q_1 = Value of $(n/4)$th observation = 2010/ 4 or 502.5th observation

This observation lies in the class interval 41–80. Therefore,

$$Q_1 = l + \frac{(n/4) - cf}{f} \times h$$

$$= 41 + \frac{502.5 - 394}{461} \times 40 = 41 + 9.41 = 50.41 \text{ acres}$$

Q_3 = Value of $(3n/4)$th observation = $(3 \times 2010)/4$ or 1507.5th observation

This observation lies in the class interval 121–160. Therefore,

$$Q_3 = l + \frac{(3n/4) - cf}{f} \times h$$

$$= 121 + \frac{1507.5 - 1246}{334} \times 40 = 121 + 31.31 = 152.31 \text{ acres}$$

Thus, the quartile deviation is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{152.31 - 50.41}{2} = 50.95 \text{ acres}$$

and

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{50.95}{202.72} = 0.251$$

Advantages and Disadvantages of Quartile Deviation

Advantages

1. Easy to calculate but useful only to evaluate variation among observed values within the middle of the data set.
2. Value is not affected by the extreme (highest and lowest) values in the data set.
3. An appropriate measure of variation for a data set having open-ended class intervals.
4. Since it is a positional measure of variation, it is useful in case of highly skewed distributions, where other measures of variation get affected by extreme values in the data set.

Disadvantages

1. Instead of all observations, the value of Q.D. is based on the middle 50 per cent values in the data set, it cannot be considered as a good measure of variation.
2. The value of Q.D. is not affected by the distribution of the individual values within the interval of the middle 50 per cent values in the data set.

Conceptual Questions 4A

1. Explain the term variation. What does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.
[Delhi Univ., MBA, 2003]

2. What are the requisites of a good measure of variation?

3. Explain how measures of central tendency and measures of variation are complementary to each other in the context of analysis of data.

4. Distinguish between absolute and relative measures of variation. Give a broad classification of the measures of variation.

5. (a) Critically examine the different methods of measuring variation.
(b) Explain with suitable examples the term 'variation'. Mention some common measures of variation and describe the one which you think is the most important. *[Delhi Univ., MBA, 2004]*

6. Explain and illustrate how the measures of variation afford a supplement to the information about frequency distribution furnished by averages.
[Delhi Univ., MBA, 2005]

7. What do you understand by 'coefficient of variation'? Discuss its importance in business problems.

Self-practice Problems 4A

4.1 The following are the prices of shares of a company from Monday to Saturday:

Days	*Price (₹)*	*Days*	*Price (₹)*
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	Saturday	250

Calculate the range and its coefficient.

4.2 The day's sales figures (in ₹) for the last 15 days at Nirula's ice-cream counter, arranged in ascending order of magnitude, are recorded as follows: 2000, 2000, 2500, 2500, 2500, 3500, 4000, 5300, 9000, 12,500, 13,500, 24,500, 27,100, 30,900, and 41,000. Determine the range and middle 50 per cent range for this sample data.

4.3 The following distribution shows the sales of the 50 largest companies for a recent year:

4.5 AVERAGE DEVIATION MEASURES

Since two measures of variation, range and quartile deviation discussed earlier do not indicate how values in a data set are scattered around a central value or disperse throughout the range, therefore it is important to measure the amount (degree) by which these values in a data set deviate from a measure of central value—usually mean or median.

To understand the nature of spread of values in the data set, two more measures of dispersion that are useful to measure the average deviation from a measure of central value—usually mean or median are:

(i) Mean Absolute Deviation or Average Deviation
(ii) Variance and Standard Deviation

4.5.1 Mean Absolute Deviation

Since average (mean) deviation of individual values in the data set from their actual arithmetic mean (A.M.) is always zero, therefore such a measure would not indicate any variation. This problem can be solved in two ways:

(i) Ignore the signs of the deviation by taking absolute value.
(ii) Square the deviations because the square of a negative number is positive.

The absolute difference between a value x_i of an observation from A.M. (or median) is always a positive number. The average value of these deviations from the A.M. (or median) is called the ***mean absolute deviation*** **(MAD)**. The MAD value is used to compare relative tendency of values in the distribution to scatter around a central value or to disperse throughout the range.

In general, the mean absolute deviation is given by

$$\text{MAD} = \frac{1}{\text{N}}\sum_{i=1}^{\text{N}} |x - \mu|, \quad \text{for a population} \tag{4-6}$$

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x - \bar{x}|, \quad \text{for a sample}$$

where | | is the sign of absolute value. That is, the plus or minus sign of deviations from the mean are ignored.

For a grouped frequency distribution, MAD is given by

$$\text{MAD} = \frac{\sum_{i=1}^{n} f_i\,|x_i - \bar{x}|}{\sum f_i} \tag{4-7}$$

While calculating MAD, the median is also considered for computing mean absolute deviation because sum of the absolute values of deviations from the median is smaller than that from any other value. However, in general, arithmetic mean is used for this purpose.

If a frequency distribution is symmetrical, then MAD taken from either mean or median is equal. Thus, the interval $\bar{x}$ ± MAD provides a range in which 57.5 per cent of the observations are included. Even if the frequency distribution is moderately skewed, the interval $\bar{x}$ ± MAD includes the same percentage of observations. This shows that more than half of the observations are scattered within one unit of the MAD around the arithmetic mean. The MAD is useful in situations where extreme deviations are likely to occur.

Coefficient of MAD

The relative measure of MAD is called the *coefficient of MAD* and is obtained by dividing the MAD by a measure of central tendency (arithmetic mean or median) used for calculating the MAD. Thus,

$$\text{Coefficient of MAD} = \frac{\text{Mean absolute deviation}}{\bar{x} \text{ or Me}} \tag{4-8}$$

If the value of relative measure is desired in percentage, then

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\bar{x} \text{ or Me}} \times 100$$

Example 4.5: The number of patients seen in the emergency ward of a hospital for a sample of 5 days in the last month was 153, 147, 151, 156 and 153. Determine the mean absolute deviation and interpret.

Solution: The mean number of patients is $\bar{x} = (153 + 147 + 151 + 156 + 153)/5 = 152$. The calculations of MAD using formula (4-6) are shown below.

Numer of Patients (x)	$x - \bar{x}$	*Absolute Deviation* $\lvert x - \bar{x} \rvert$
153	153 – 152 = 1	1
147	147 – 152 = –5	5
151	151 – 152 = –1	1
156	156 – 152 = 4	4
153	153 – 152 = 1	1
		12

$$\text{MAD} = \frac{1}{n}\sum \lvert x - \bar{x} \rvert = \frac{12}{5} = 2.4 \cong 3 \text{ patients (approx)}$$

The mean absolute deviation is 3 patients per day. The deviation in the number of patients falls in the interval 152 ± 3 patients per day.

Example 4.6: Calculate the mean absolute deviation and its coefficient from median for the following data

Year	*Sales (₹ thousand)*	
	Product A	*Product B*
2006	23	36
2007	41	39
2008	29	36
2009	53	31
2010	38	47

Solution: The median sales of the two products A and B is 38 and 36, respectively. The calculations of MAD in both the cases are shown in Table 4.2.

Table 4.2 Calculations of MAD

Product A		*Product B*	
Sales (x)	$\lvert x - Me \rvert = \lvert x - 38 \rvert$	*Sales (x)*	$\lvert x - Me \rvert = \lvert x - 36 \rvert$
23	15	31	5
29	9	36	0
38	0	36	0
41	3	39	3
53	15	47	11
$n = 5$	$\Sigma \lvert x - \text{Me} \rvert = 42$	$n = 5$	$\Sigma \lvert x - \text{Me} \rvert = 19$

Product A:
$$\text{MAD} = \frac{1}{n}\sum \lvert x - \text{Me} \rvert = \frac{42}{5} = 8.4$$

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\text{Me}} = \frac{8.4}{38} = 0.221$$

Product B: $$\text{MAD} = \frac{1}{n}\sum |x - \text{Me}| = \frac{19}{5} = 3.8$$

$$\text{Coefficient of MAD} = \frac{\text{MAD}}{\text{Me}} = \frac{3.8}{36} = 0.106$$

Example 4.7: Find the mean absolute deviation from mean for the following frequency distribution of sales (₹ in thousand) in a co-operative store.

Sales	:	50–100	100–150	150–200	200–250	250–300	300–350
Number of days	:	11	23	44	19	8	7

Solution: The mean absolute deviation can be calculated by using the formula (4-6) for A.M. ($\bar{x}$). The calculations for MAD are shown in Table 4.3. Let, assumed mean, A = 175.

Table 4.3 Calculations for MAD

Sales (*Rs*)	*Mid-Value* (*m*)	*Frequency* (*f*)	*d* = (*m* −*175*)/*50*	*fd*	$\lvert x - \bar{x} \rvert = \lvert m - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
50 – 100	75	11	−2	−22	104.91	1154.01
100 – 150	125	23	−1	−23	54.91	1262.93
150 – 200	175 ← A	44	0	0	4.91	216.04
200 – 250	225	19	1	19	45.09	856.71
250 – 300	275	8	2	16	95.09	760.72
300 – 350	325	7	3	21	145.09	1015.63
		112		11		5266.04

$$\bar{x} = \text{A} + \left\{\frac{1}{n}\sum fd\right\} \times h = 175 + \frac{11}{112} \times 50 = ₹179.91 \text{ per day}$$

$$\text{MAD} = \frac{1}{n}\sum f\,|x - \bar{x}| = \frac{5266.04}{112} = ₹47.01$$

Thus, the average sales are ₹1,79,910 per day and the mean absolute deviation of sales is ₹47,010 per day.

Advantages and Disadvantages of MAD

Advantages

1. The calculation of MAD is based on all observations in the distribution and shows the dispersion of values around the measure of central tendency.
2. While calculating MAD, equal weightage is given to each observed value to indicate how far each observation lies from either the mean or median.
3. Average deviation from arithmetic mean is always zero in any data set. In MAD this problem is taken care by using absolute values to eliminate the negative signs.

Disadvantages

1. While calculating MAD, the algebraic signs are ignored. If the signs are not ignored, then sum of the deviations taken from arithmetic mean will be zero and close to zero when deviations are taken from median.
2. The value of MAD is considered to be best when deviations are taken from median. But median does not provide a satisfactory result in case the amount of variation is more in a data set.

4.5.2 Variance and Standard Deviation

While computing absolute value of each deviation from arithmetic mean, another way to ignore sign of negative deviations from mean is to square such values. The sum of all such squared deviations is then divided by the number of observations in the data set. A value so obtained is called **population variance** denoted by σ^2 (a lower-case Greek letter sigma). It is usually referred to as 'sigma squared'. Symbolically, it is written as

Variance: A measure of variability based on the squared deviations of the observed values in the data set about arithmetic mean.

$$\text{Population variance, } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad (4\text{-}9)$$

$$= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - (\mu)^2$$

(Deviation is taken from actual population A.M.)

$$= \frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2 \text{ (Deviation is taken from assumed mean, A)}$$

where $d = x - A$ and A is any constant (also called assumed A.M.)

Since σ^2 is the average (or mean) of squared deviations from arithmetic mean, it is also called the ***mean square average.***

The population variance is used to measure variation among the values of observations in a population. However, in almost all applications of statistics, the data being analysed is a sample data. Thus, sample variance is determined to estimate population variance, σ^2.

If the *sum of the squared deviations* about a sample mean $\bar{x}$ in Eq. (4-9) is divided by n (sample size), then invariably the estimated value of σ^2 is lower than its actual value. Such a difference in two values is called *bias*. However, this *bias* in the estimation of population variance from a sample variance can be removed by dividing the sum of the squared deviations between the sample mean and each value in the population by $n - 1$ rather than by n. The *unbiased sample variance* denoted by s^2 is defined as follows:

$$\text{Sample variance, } s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum x^2}{n-1} - \frac{n\,\bar{x}^2}{n-1} = \frac{\Sigma x^2}{n-1} - \frac{(\Sigma x)^2}{n(n-1)}. \qquad (4\text{-}10)$$

Standard Deviation: A measure of variability computed by taking the positive square root of the variance.

The numerator $\sum(x-\bar{x})^2$ in Eq. (4-10) is called the *total sum of squares*. This quantity measures the total variation among values in a data set (whereas the variance measures only the *average variation*). The larger the value of $\Sigma(x - \bar{x})^2$, the greater the variation among the values in a data set.

Standard Deviation

The numerical value of population or a sample variance is difficult to interpret because it is expressed in square units. To reach an interpretable measure of variance expressed in the units of original data, we take a positive square root of the variance, called ***standard deviation or root-mean square deviation***. The standard deviation of population and sample is denoted by σ and s, respectively.

(a) *Ungrouped Data*

$$\text{Population standard deviation, } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum(x-\mu)^2} = \sqrt{\frac{1}{N}\sum x^2 - (\mu)^2}$$

$$= \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$$

$$\text{Sample standard deviation, } s = \sqrt{\frac{\Sigma x^2}{n-1} - \frac{n\bar{x}^2}{n-1}}\text{; where } n = \text{sample size}$$

(b) *Grouped Data*

$$\text{Population standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h$$

where f is the frequency of each class interval; N is the total number of observations (or elements) in the population; h is the width of class interval; m is mid-value of each class interval and $d = (m - A) / h$, where A is any constant (also called assumed A.M.)

$$\text{Sample standard deviation, } s = \sqrt{s^2} = \sqrt{\frac{\sum f(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum fx^2}{n-1} - \frac{(\sum fx)^2}{n(n-1)}} \qquad (4\text{-}11)$$

Remarks: 1. For any data set, MAD is always less than the σ because MAD is less sensitive to the extreme observations. Thus, when a data contains few outliers, the MAD provides a more realistic measure of variation than σ. However, σ is often used in statistical applications because formula is capable of algebraic treatment.

2. When sample size (n) becomes very large, (n – 1) becomes irrelevant.

Advantages and Disadvantages of Standard Deviation

Advantages

1. The value of standard deviation is based on every observation in a set of data. The formula of standard deviation is capable of algebraic treatment and is less affected by fluctuations of sample size as compared to other measures of variation.
2. It is possible to calculate the combined standard deviation of two or more sets of data.
3. The area under the symmetric curve of a frequency distribution is expressed in terms of standard deviation and population mean.
4. Standard deviation is used for comparing skewness, correlation, and so on, and also widely used in sampling theory.

Disadvantages

1. Calculations of standard deviation are slightly difficult as compared to other measures of variation.
2. Since for calculating S.D., the deviations from the arithmetic mean are squared, therefore large deviations when squared are proportionately more than small deviations. For example, the deviations 2 and 10 are in the ratio of 1 : 5 but their squares 4 and 100 are in the ratio of 1 : 25.

Example 4.8: The wholesale prices of a commodity for seven consecutive days in a month are as follows:

Days	:	1	2	3	4	5	6	7
Commodity price/quintal	:	240	260	270	245	255	286	264

Calculate the variance and standard deviation.

Solution: The computations for variance and standard deviation from actual arithmetic mean, $\bar{x}$ are shown in Table 4.4.

Table 4.4 Computations of Variance and Standard Deviation with Actual Mean

Observation (x)	$x - \bar{x} = x - 260$	$(x - \bar{x})^2$
240	−20	400
260	0	0
270	10	100
245	−15	225
255	−5	25
286	26	676
264	4	16
1820		1442

$$\bar{x} = \frac{1}{n}\sum x = \frac{1}{7}(1820) = 260$$

$$\text{Variance } \sigma^2 = \frac{1}{n}\sum(x-\bar{x})^2 = \frac{1}{7}(1442) = 206$$

$$\text{Standard deviation } \sigma = \sqrt{\sigma^2} = \sqrt{206} = 14.352$$

If deviation is taken from an assumed A.M. = 255 instead of actual A.M. = 260, then calculations for standard deviation are shown in Table 4.5.

Table 4.5 Computations of Standard Deviation with Assumed Mean

Observation (x)	$d = x - A = x - 255$	d^2
240	−15	225
260	5	25
270	15	225
245	−10	100
$\boxed{255}$ ← A	0	0
286	31	961
264	9	81
	35	1617

$$\text{Standard deviation } \sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{1617}{7} - \left(\frac{35}{7}\right)^2}$$

$$= \sqrt{231 - 25} = \sqrt{206} = 14.352$$

This result is same as shown in Table 4.4.

Remark: When actual A.M. is not a whole number, assumed A.M. method should be used to reduce the computation time.

Example 4.9: A study of 100 engineering companies gives the following information

Profit (₹ in crore)	:	0–10	10–20	20–30	30–40	40–50	50–60
Number of companies	:	8	12	20	30	20	10

Calculate the standard deviation of the profit earned.

Solution: Let assumed mean, A be 35. Calculations for standard deviation are shown in Table 4.6.

Table 4.6 Calculations of Standard Deviation

Profit (*₹ in crore*)	*Mid-value* (*m*)	$d = \frac{m-A}{h} = \frac{m-35}{10}$	*Number of Companies* (*f*)	*fd*	fd^2
0–10	5	−3	8	−24	72
10–20	15	−2	12	−24	48
20–30	25	−1	20	−20	20
30–40	$\boxed{35}$ ← A	0	30	0	0
40–50	45	1	20	20	20
50–60	55	2	10	20	40
			100	−28	200

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times h$$

$$= \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 10 = \sqrt{2 - 0.078} \times 10 = 13.863$$

Example 4.10: Mr Gupta, a retired government servant, is considering investing his money in two proposals. He wants to choose the one that has higher average net present value and lower standard deviation. The relevant data are given below. Can you help him in choosing the proposal?

Proposal A:

Net Present Value (NPV)	*Chance of the Possible Outcome of NPV*
1559	0.30
5662	0.40
9175	0.30

Proposal B:

Net Present Value (NPV)	*Chance of the Possible Outcome of NPV*
– 10,050	0.30
5,812	0.40
20,584	0.30

Solution: The expected (average) net present value for both the proposals is:

Proposal A: Expected NPV $= 1559 \times 0.30 + 5662 \times 0.40 + 9175 \times 0.30$
$= 467.7 + 2264.8 + 2752.5 =$ ₹5485

Proposal B: Expected NPV $= -\ 10{,}050 \times 0.30 + 5812 \times 0.40 + 20{,}584 \times 0.30$
$= -\ 3015 + 2324.8 + 6175.2 =$ ₹5485

Since the expected NPV in both the cases is same, Mr. Gupta would like to choose less risky proposal. For this, we have to calculate the standard deviation in both the cases.

Standard deviation for proposal A:

NPV(x_i)	*Expected NPV* ($\bar{x}$)	$x - \bar{x}$	*Probability of NPV* (f)	$f(x - \bar{x})^2$
1559	5485	–3926	0.30	46,24,042.8
5662	5485	177	0.40	12,531.6
9175	5485	3690	0.30	40,84,830.0
			1.00	87,21,404.4

$$s_A = \sqrt{\frac{\sum f(x - x)^2}{N}} = \sqrt{87{,}21{,}404.4} = ₹2953.20$$

Standard deviation for proposal B:

NPV(x_i)	*Expected NPV* ($\bar{x}$)	$x - \bar{x}$	*Probability of NPV* (f)	$f(x - \bar{x})^2$
– 10,050	5485	– 15,535	0.30	7,24,00,867.5
5812	5485	327	0.40	42,771.6
20,584	5485	15,099	0.30	6,83,93,940
			1.00	14,08,37,579

$$s_B = \sqrt{\frac{\sum f(x - \bar{x})^2}{N}} = \sqrt{14{,}08{,}37{,}579} = ₹11{,}867.50$$

Since $s_A < s_B$, therefore proposal A indicates uniform net profit and hence may be chosen.

4.5.3 Mathematical Properties of Standard Deviation

1. ***Combined standard deviation***: The combined standard deviation, σ_{12} of two sets of data containing n_1 and n_2 observations with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2, respectively, is given by

$$\sigma_{12} = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$$

where $d_1 = \bar{x}_{12} - \bar{x}_1 \,;\; d_2 = \bar{x}_{12} - \bar{x}_2$

and $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ (combined arithmetic mean)

This formula can also be extended to compute the standard deviation of more than two sets of data.

2. ***Standard deviation of natural numbers***: The standard deviation of the first n natural numbers is given by

$$\sigma = \sqrt{\frac{1}{12}(n^2 - 1)}$$

For example, the standard deviation of the first 100 (i.e., from 1 to 100) natural numbers will be

$$\sigma = \sqrt{\frac{1}{12}(100^2 - 1)} = \sqrt{\frac{1}{12}(9999)} = \sqrt{833.25} = 28.86$$

Example 4.11: For a group of 50 male workers, the mean and standard deviation of their monthly wages are ₹6300 and ₹900, respectively. For a group of 40 female workers, these are ₹5400 and ₹600, respectively. Find the standard deviation of monthly wages for the combined group of workers. *[Delhi Univ., MBA, 2004]*

Solution: Given that, Male workers : $n_1 = 50,\ \bar{x}_1 = 6300,\ \sigma_1 = 900$

Female workers : $n_2 = 40,\ \bar{x}_2 = 5400,\ \sigma_2 = 600$

Then, Combined mean, $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \dfrac{50 \times 6300 + 40 \times 5400}{50 + 40} = 5{,}900$

and Combined Standard Deviation

$$\sigma_{12} = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$$

$$= \sqrt{\frac{50(8{,}10{,}000 + 1{,}60{,}000) + 40(3{,}60{,}000 + 2{,}50{,}000)}{50 + 40}} = ₹900$$

where $d_1 = \bar{x}_{12} - \bar{x}_1 = 5900 - 6300 = -400$ and $d_2 = \bar{x}_{12} - \bar{x}_2 = 5900 - 5400 = 500.$

Example 4.12: A study of the age of 100 persons grouped into intervals 20–22, 22–24, 24–26, ... revealed the mean age and standard deviation to be 32.02 and 13.18, respectively. While checking, it was discovered that the observation 57 was misread as 27. Calculate the correct mean age and standard deviation. *[Delhi Univ., MBA, 2007]*

Solution: From the data given in the problem, we have $\bar{x} = 32.02$, $\sigma = 13.18$ and N = 100. We know that

$$\bar{x} = \frac{1}{n}\sum fx \quad \text{or} \quad \Sigma fx = n \times \bar{x} = 100 \times 32.02 = 3202$$

and

$$\sigma^2 = \frac{1}{n}\sum fx^2 - (\bar{x})^2 \text{ or } \Sigma fx^2 = n[\sigma^2 + (\bar{x})^2] = 100[(13.18)^2 + (32.02)^2]$$

$$= 100[173.71 + 1025.28] = 100 \times 1198.99 = 1{,}19{,}899$$

On substituting the correct observation, we get

$$\Sigma fx = 3202 - 27 + 57 = 3232.$$

Also $\Sigma fx^2 = 1,19,899 - (27)^2 + (57)^2 = 1,19,899 - 729 + 3248 = 1,22,419$

Thus, Correct A.M., $\bar{x} = \frac{1}{n}\sum fx = \frac{1}{100}(3232) = 32.32.$

and Correct variance, $\sigma^2 = \frac{1}{n}\sum fx^2 - (\bar{x})^2 = \frac{1}{100}(1,22,419) - (32.32)^2$

$$= 1224.19 - 1044.58 = 179.61$$

or Correct standard deviation, $\sigma = \sqrt{\sigma^2} = \sqrt{179.61} = 13.402.$

Example 4.13: The mean of 5 observations is 15 and the variance is 9. If two more observations having values –3 and 10 are combined with these 5 observations, what will be the new mean and variance of 7 observations?

Solution: From the data of the problem, we have $\bar{x} = 15, s^2 = 9$ and $n = 5$. We know that

$$\bar{x} = \frac{1}{n}\sum x \quad \text{or} \quad \sum x = n \times \bar{x} = 5 \times 15 = 75$$

If two more observations having values –3 and 10 are added to the existing 5 observations, then after adding these 6th and 7th observations, we get

$$\sum x = 75 - 3 + 10 = 82$$

Thus, New A.M., $\bar{x} = \frac{1}{n}\sum x = \frac{1}{7}(82) = 11.71$

$$\text{Variance, } s^2 = \frac{1}{n}\sum x^2 - (\bar{x})^2$$

$$9 = \frac{1}{n}\sum x^2 - (15)^2$$

or $\sum x^2 = 1170$

On adding two more observations: –3 and 10, we get

$$\sum x^2 = 1170 + (-3)^2 + (10)^2 = 1279$$

$$\text{Variance, } s^2 = \frac{1}{n}\sum x^2 - (\bar{x})^2 = \frac{1}{7}(1279) - (11.71)^2 = 45.59$$

Hence, the new mean and variance of 7 observations is 11.71 and 45.59, respectively.

4.5.4 Chebyshev's Theorem

Standard deviation measures the variation among observations in a data set. If the standard deviation value is small, then values in the data set cluster close to the arithmetic mean. Conversely, a large standard deviation value indicates that the values are scattered more widely around arithmetic mean. P. L. Chebyshev (1821–1894) a Russian mathematician, developed a result called ***Chebyshev's theorem*** to indicate proportion of values in the data set that fall within a specified number of standard deviation from the mean value. The theorem states that:

Chebyshev's Theorem: A statement about the proportion of observations that must lie within σ, 2σ, and 3σ deviations from the mean (population or sample distribution).

For any set of data (population or sample) and any constant z greater than 1 (but need not be an integer), the proportion of the values that lie within z standard deviations on either side of the mean is at least $\{1 - (1/z^2)\}$. *That is*

$$\text{RF}\,[\,|x - \mu| \le z\sigma] \ge 1 - \frac{1}{z^2}$$

where RF is the relative frequency of a distribution.

$$z = \frac{x-\mu}{\sigma}$$ ← population standardized score, i.e., number of standard deviations a value, x is away from the mean μ (sample or population)

$$= \frac{x-\bar{x}}{s}$$ ← sample standardized score

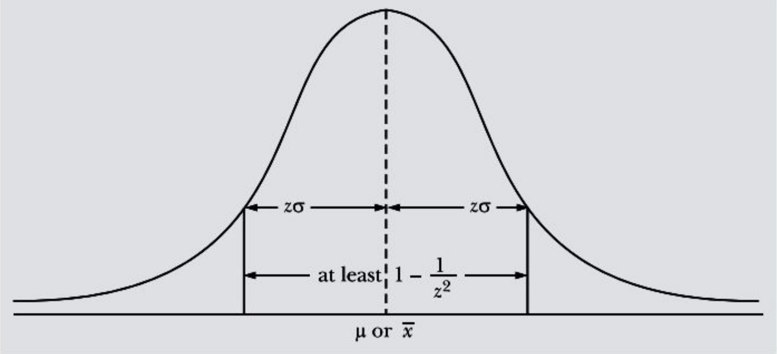


Figure 4.5
Chebyshev's Theorem

For a symmetrical, bell-shaped distribution as shown in Fig. 4.5. Chebyshev's theorem indicates percentage of values that *approximately* fall within z standard deviations. The relationship among mean, standard deviation and the set of values is called ***empirical* (or *normal) rule*.**

Illustration The theorem is applicable to any data set regardless of the shape of the frequency distribution of values. For example, assume that the marks obtained by 100 students in business statistics had an A.M., $\bar{x}$ = 70 per cent and standard deviation, σ = 10 per cent. Then number of students who obtained marks between 50 and 85 will be determined as follows:

(a) Since, $z = (50 - 70)/10 = -2$, 50 marks fall 2 standard deviations below the mean,

(b) Since, $z = (85 - 70)/10 = 1.5$, 85 marks fall 1.5 standard deviations above the mean.

Applying the Chebyshev's theorem with $z = 2.0$, we have

$$\left(1 - \frac{1}{z^2}\right) = \left[1 - \frac{1}{(2.0)^2}\right] = 0.75$$

This indicates that at least 75 per cent of the students must have obtained marks between 50 and 85.

Empirical Rule

For symmetrical, bell-shaped frequency distribution (also called normal curve), the range within which a given percentage of values of the distribution are likely to fall within a specified number of standard deviations of the population mean, μ is determined as follows:

$\mu \pm \sigma$ covers approximately 68.27 per cent of values in the data set.

$\mu \pm 2\sigma$ covers approximately 95.45 per cent of values in the data set.

$\mu \pm 3\sigma$ covers approximately 99.73 per cent of values in the data set.

These ranges are illustrated in Fig. 4.5.

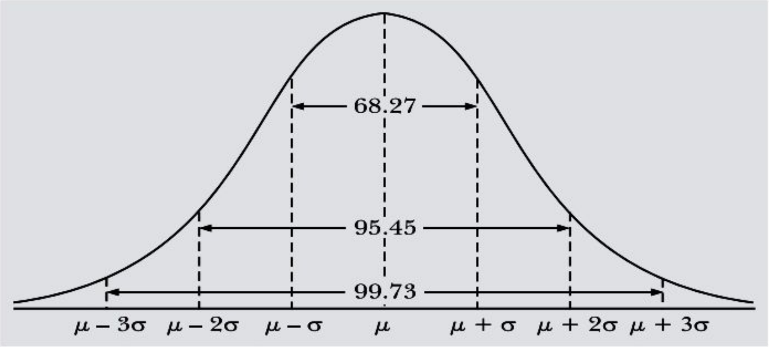


Figure 4.6
Area Under Normal Curve

For a symmetrical and bell-shaped distribution, relationships among three measures of variation are given in Table 4.10.

Table 4.7 Relationship Among Measures of Variation

Measures of Variation	*Percentage of Values Scatter Around the Mean Value, μ*			*Size of Measure of Variation to Standard Deviation at*
	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$	
Q.D.	50.00	82.30	95.70	0.6748
MAD	57.50	88.90	98.30	0.7979
S.D.	68.27	95.45	99.73	1.0000

Relationship Between Different Measures of Variation

(a) Quartile deviation (Q.D.) $= \frac{2}{3}\sigma$

Mean absolute deviation (MAD) $= \frac{4}{5}\sigma$

(b) Quartile deviation $= \frac{5}{6}$MAD

Standard deviation $= \frac{5}{4}$MAD or $\frac{3}{2}$Q.D.

(c) Mean absolute deviation $= \frac{6}{5}$Q.D.

These relationships are applicable only to symmetrical distributions.

Example 4.14: Suppose you are in charge of rationing in a state affected by food shortage. The following reports arrive from a local investigator:

Daily caloric value of food available per adult during current period:

Area	*Mean*	*Standard Deviation*
A	2500	400
B	2000	200

The estimated requirement of an adult is taken as 2800 calories daily and the absolute minimum is 1350. Comment on the reported figures and determine which area in your opinion, need more urgent attention.

Solution: Taking into consideration the entire population of the two areas, we have

Area A: $\mu + 3\sigma = 2500 + 3 \times 400 = 3700$ calories
$\mu - 3\sigma = 2500 - 3 \times 400 = 1300$ calories

This calculation shows that adults are taking only 1300 calories, which is much less than the absolute minimum requirement of 1350 calories.

Area B: $\mu + 3\sigma = 2000 + 3 \times 200 = 2600$ calories
$\mu - 3\sigma = 2000 - 3 \times 200 = 1400$ calories

This calculation shows that adults are taking sufficient amount of calories as per requirement of daily calorific need. Hence, area A needs more urgent attention.

Example 4.15: The following data give the number of passengers travelling by airplane from one city to another in one week.

115 122 129 113 119 124 132 120 110 116

Calculate the mean and standard deviation and determine the percentage of class that lie between (i) $\mu \pm \sigma$, (ii) $\mu \pm 2\sigma$ and (iii) $\mu \pm 3\sigma$. What percentage of cases lies outside these limits?

Solution: The calculations for mean and standard deviation are shown in Table 4.8.

Table 4.8 Calculations of Mean and Standard Deviation

x	$x - \bar{x}$	$(x - \bar{x})^2$
115	−5	25
122	2	4
129	9	81
113	−7	49
119	−1	1
124	4	16
132	12	144
120	0	0
110	−10	100
116	−4	16
1200	0	436

$$\mu = \frac{1}{n}\sum x = \frac{1200}{10} = 120 \text{ and } \sigma^2 = \frac{1}{n}\sum(x-\bar{x})^2 = \frac{436}{10} = 43.6$$

Therefore, $\sigma = \sqrt{\sigma^2} = \sqrt{43.6} = 6.60$

The percentage of cases that lies between a given limit is as follows:

Interval	*Values within Interval*	*Percentage of Population*	*Percentage Falling Outside*
$\mu \pm \sigma = 120 \pm 6.60$ $= 113.4$ and 126.6	113, 115, 116, 119 120, 122, 124	70%	30%
$\mu \pm 2\sigma = 120 \pm 2(6.60)$ $= 106.80$ and 133.20	110, 113, 115, 116, 119 120, 122, 124, 129, 132	100%	nil

Example 4.16: A collar manufacturer is considering the production of a new collar to attract young men. Thus, following statistics of neck circumference are available based on measurement of a typical group of the college students:

Mid value (in inches):	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0
Number of students :	2	16	36	60	76	37	18	3	2

Compute the standard deviation and use the criterion $\bar{x} \pm 3\sigma$, where σ is the standard deviation and $\bar{x}$ is the arithmetic mean to determine the largest and smallest size of the collar he should make in order to meet the needs of practically all the customers bearing in mind that collar are worn on average half inch longer than neck size.

Solution: Calculations for mean and standard deviation in order to determine the range of collar size to meet the needs of customers are shown in Table 4.9.

Table 4.9 Calculations for Mean and Standard Deviation

Mid-value (in inches)	*Number of students*	$\frac{x-A}{h} = \frac{x-14}{0.5}$	*fd*	*fd*^{2}
12.0	2	−4	−8	32
12.5	16	−3	−48	144
13.0	36	−2	−72	144
13.5	60	−1	−60	60
14.0 ← A	76	0	0	0
14.5	37	1	37	37
15.0	18	2	36	72
15.5	3	3	9	27
16.0	2	4	8	32
	N = 250		− 98	548

$$\text{Mean, } \bar{x} = \text{A} + \frac{\sum fd}{n} \times h = 14.0 - \frac{98}{250} \times 0.5 = 14.0 - 0.195 = 13.805$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum fd^2}{n} - \left(\frac{\sum fd}{n}\right)^2} \times h = \sqrt{\frac{548}{250} - \left(\frac{-98}{250}\right)^2} \times 0.5$$

$$= \sqrt{2.192 - 0.153} \times 0.5 = 1.427 \times 0.5 = 0.7135$$

Largest and smallest neck size = $\bar{x} \pm 3\sigma = 13.805 \pm 3 \times 0.173 = 11.666$ and 15.944.

Since all the customers are to wear collar half inch longer than their neck size, 0.5 is to be added to the neck size range given above. The new range then becomes

(11.666 + 0.5) and (15.944 + 0.5) or 12.2 and 16.4 inches (approx).

Example 4.17: A welfare organization introduced an education scholarship scheme for school going children of a backward village. The rates of scholarship were fixed as given below:

Age Group (Years)	*Amount of Scholarship per Month (₹)*
5–7	300
8–10	400
11–13	500
14–16	600
17–19	700

The age of 30 school children is: 11, 8, 10, 5, 7, 12, 7, 17, 5, 13, 9, 8, 10, 15, 7, 12, 6, 7, 8, 11, 14, 18, 6, 13, 9, 10, 6, 15, 3, 5 years, respectively. Calculate mean and standard deviation of monthly scholarship. Find out the total monthly scholarship amount being paid to the students. *[IGNOU, MBA, 2002]*

Solution: The number of students in the age group from 5–7 to 17–19 are calculated as shown in Table 4.10:

Table 4.10

Age Group (Years)	*Tally Bars*	*Number of Students*
5–7	𝍸 𝍸	10
8–10	𝍸 |||	8
11–13	𝍸 ||	7
14–16	|||	3
17–19	||	2
		30

The calculations for mean and standard deviation are shown in Table 4.11.

Table 4.11 Calculations for Mean and Standard Deviation

Age Group (Years)	*Number of Students (f)*	*Mid-value (m)*	$d = \frac{m-A}{h} = \frac{m-12}{3}$	*fd*	*fd*^{2}
5 – 7	10	6	– 2	– 20	40
8 – 10	8	9	– 1	– 8	8
11 – 13	7	$A \rightarrow \textcircled{12}$	0	0	0
14 – 16	3	15	1	3	3
17 – 19	2	18	2	4	8
	30			– 21	59

$$\text{Mean} = A + \left\{\frac{1}{n}\Sigma fd\right\} \times h = 12 - \frac{21}{30} \times 3 = 12 - 2.1 = 9.9$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2} \times h = \sqrt{\frac{59}{30} - \left(\frac{-21}{30}\right)^2} \times 3$$

$$= \sqrt{1.967 - 0.49} \times 3 = 1.2153 \times 3 = 3.6459$$

Calculations for monthly scholarship paid to 30 students are shown in Table 4.12.

Table 4.12 Calculations for Monthly Scholarship

Number of Students	*Amount of Scholarship per Month (₹)*	*Total Monthly Scholarship (₹)*
10	300	3000
8	400	3200
7	500	3500
3	600	1800
2	700	1400
		12,900

Example 4.18: The breaking strength of 80 'test pieces' of a certain alloy is given in the following table, the unit being given to the nearest thousand grams per square inch:

Breaking Strength	*Number of Pieces*
44–46	3
46–48	24
48–50	27
50–52	21
52–54	5

Calculate the average breaking strength of the alloy and the standard deviation. Calculate the percentage of observations lying between $\overline{x} \pm 2\sigma$. *[Vikram Univ., MBA, 2005]*

Solution: The calculations for mean and standard deviation are shown in Table 4.13.

Table 4.13 Calculations for Mean and Standard Deviation

Breaking Strength	*Number of Pieces(f)*	*Mid-value (m)*	*d = (m – A)/h* *= (m – 49/2*	*fd*	*fd*^{2}
44–46	3	45	–2	–6	12
46–48	24	47	–1	–24	24
48–50	27	A → 49	0	0	0
50–52	21	51	1	21	21
52–54	5	53	2	10	20
	80			1	77

Mean, $\bar{x}$ = A + = 49 + × 2 = 49.025

Standard deviation,
$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times h = \sqrt{\frac{77}{80} - \left(\frac{1}{80}\right)^2} \times 2$$
$$= \sqrt{0.9625 - 0.000} \times 2 = 0.9810 \times 2 = 1.962$$

Breaking strength of pieces in the range, $\bar{x} \pm 2\sigma$ is

$$\bar{x} \pm 2\sigma = 49.025 \pm 2 \times 1.962$$
$$= 45.103 \text{ and } 52.949 = 45 \text{ and } 53 \text{ (approx.)}$$

To calculate the percentage of observations that fall in range $\bar{x} \pm 2\sigma$, it is assumed that the equal number of observations is spread within lower and upper limit of each class interval. Since 45 is the mid-point of the class interval 44–46 with the frequency 3, therefore there are 1.5 frequencies at 45. Similarly, at 53 the frequency would be 2.5. Hence, the total number of observations (frequencies) between 45 and 53 are 1.5 + 24 + 27 + 21 + 2.5 = 76. Thus, percentage of observations lying within range, $\bar{x} \pm 2\sigma$ would be (76/80) × 100 = 95 per cent.

4.5.5 Coefficient of Variation

A relative measure called the **coefficient of variation** (CV) developed by Karl Pearson is very useful measure for (i) comparing two or more data sets expressed in different units of measurement, and (ii) comparing data sets that are in same unit of measurement but the mean values of data sets are not same.

The coefficient of variation (CV) that measures the standard deviation relative to the mean in percentages is computed as follows:

Coefficient of Variation: A measure of relative variability computed by dividing the standard deviation by the mean, then multiplying by 100.

$$\text{Coefficient of variation (CV)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

Multiplying by 100 converts the decimal to a percent. The lower value of CV indicates uniformity (or consistency) among values in any data set.

Example 4.19: The weekly sales of two products A and B were recorded as given below:

Product A	:	59	75	27	63	27	28	56
Product B	:	150	200	125	310	330	250	225

Find out which of the two shows greater fluctuation in sales.

Solution: Calculating coefficient of variation for both the products to compare fluctuation in their sales.

Product A: Let A = 56 be the assumed mean of sales for product A.

Table 4.14 Calculations of the Mean and Standard Deviation

Sales (x)	*Frequency (f)*	$d = x - A$ $= x - 56$	*fd*	fd^2
27	2	−29	−58	1682
28	1	−28	−28	784
56 ← A	1	0	0	0
59	1	3	3	9
63	1	7	7	49
75	1	19	19	361
	7		−57	2885

$$\bar{x} = A + \frac{1}{n}\sum fd = 56 - \frac{57}{7} = 47.86$$

$$s_A^2 = \frac{1}{n}\sum fd^2 - \left(\frac{1}{n}\sum fd\right)^2 = \frac{2885}{7} - \left(-\frac{57}{7}\right)^2$$

$$= 412.14 - 66.30 = 345.84$$

$$s_A = \sqrt{345.84} = 18.59$$

Then $\text{CV (A)} = \frac{s_A}{\bar{x}} \times 100 = \frac{18.59}{47.86} \times 100 = 38.84$ per cent

Product B: Let A = 225 be the assumed mean of sales for product B.

Table 4.15 Calculations of Mean and Standard Deviation

Sales (x)	*Frequency (f)*	$d = x - A$ $= x - 225$	*fd*	fd^2
125	1	−100	−100	10,000
150	1	−75	−75	5625
200	1	−25	−25	625
225 ← A	1	0	0	0
250	1	25	25	625
310	1	85	85	7225
330	1	105	105	11,025
	7		15	35,125

$$\bar{x} = A + \frac{1}{n}\sum fd = 225 + \frac{15}{7} = 227.14$$

$$s_B^2 = \frac{1}{n}\sum fd^2 - \left(\frac{1}{n}\sum fd\right)^2 = \frac{35{,}125}{7} - \left(\frac{15}{7}\right)^2 = 5017.85 - 4.59 = 5013.26$$

or $s_B = \sqrt{5013.26} = 70.80$

Then $\text{CV(B)} = \frac{s_B}{\bar{x}} \times 100 = \frac{70.80}{227.14} \times 100 = 31.17$ per cent

Since the coefficient variation for product A is more than that of product B, therefore the sales fluctuation in case of product A is higher.

Example 4.20: From the analysis of monthly wages paid to employees in two service organizations X and Y, the following results were obtained:

	Organization X	*Organization Y*
Number of wage-earners	550	650
Average monthly wages	5000	4500
Variance of the distribution of wages	900	1600

(a) Which organization pays a larger amount as monthly wages?

(b) In which organization is there greater variability in individual wages of all the wage earners taken together?

Solution: (a) Comparing the total wages to find out which organization X or Y pays larger amount of monthly wages:

Total wage bill paid monthly by X and Y is

$$X : n_1 \times \overline{x}_1 = 550 \times 5000 = ₹27{,}50{,}000$$
$$Y : n_2 \times \overline{x}_2 = 650 \times 4500 = ₹29{,}25{,}000$$

Organization Y pays a larger amount as monthly wages as compared to organization X.

(b) For calculating the combined variation, first calculating the combined mean as follows:

$$\overline{x}_{12} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2} = \frac{27{,}50{,}000 + 29{,}25{,}000}{1200} = ₹4729.166$$

$$\sigma_{12} = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_1^2\right)}{n_1 + n_2}}$$

$$= \sqrt{\frac{550(900 + 73{,}351.05) + 650(1600 + 52{,}517.05)}{550 + 650}}$$

$$= \sqrt{\frac{4{,}08{,}38{,}080.55 + 3{,}51{,}76{,}082.50}{1200}}$$

$$= \sqrt{63345.13} = 251.68$$

where $d_1 = \overline{x}_{12} - \overline{x}_1 = 4729.166 - 5000 = -270.834$

$d_2 = \overline{x}_{12} - \overline{x}_2 = 4729.166 - 4500 = 229.166$

Conceptual Questions 4B

8. What purpose does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.

9. What do you understand by 'coefficient of variation'? Discuss its importance in business problems. [*UP Tech. Univ., MBA, 2003*]

10. When is the variance equal to the standard deviation? Under what circumstances can variance be less than the standard deviation? Explain.

11. (a) Explain and illustrate how the measures of variation afford a supplement to the information about frequency distribution furnished by averages. [*Delhi Univ., MBA, 2004*]

(b) Describe various methods of measuring variation. Which of these do you consider as the best and why?

12. Explain the advantages of standard deviation as a measure of variation over range and the average deviation. Under what circumstances will the variance of a variable be zero?

13. Comment on the comparative merits and demerits of measures of variation.

14. Explain the term 'variation'. What purpose does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation. [*Delhi Univ., MBA, 2008*]

15. Describe the various methods of measuring variation along with their respective merits and demerits. [*Delhi Univ., MBA, 2008*]

16. It has been said that the lesser the variability that exists, the more an average is representative of a set of data. Comment.

17. (a) What information is provided by variance or standard deviation?

(b) What additional information about a set of data is provided by a measure of variability that is not obtained from an average?

18. What advantages are associated with variance and standard deviation relative to range as the measure of variability?

19. Suppose you read a published statement that the average amount of food consumption in this country is adequate; the overall conclusion based upon the statement is that everyone is properly fed. Criticize the conclusion in terms of the concept of variability as it relates to the use of averages.[*Delhi Univ., MBA, 2006*]

20. The Vice-President, Sales has been studying records regarding the performance of his sales representatives. He has noticed that in the last 2 years, the average level of sales per representative has remained the same, while the distribution of the sales levels has widened. The sales levels from this period have significantly larger variations from the mean than in any of the previous 2 year periods for which he has records. What conclusions might be drawn from these observations? [*Delhi Univ., MBA, 2009*]

21. Explain Chebyshev's theorem which provides an approximation to the spread of a set of observations on either side of the mean.

22. Two economists are studying fluctuations in the price of gold. One is examining the period of 1998–2002. The other is examining the period of 1995–1999. What differences would you expect to find in the variability of their data?

23. How would you reply to the following statement: 'Variability is not an important factor because even though the outcome is more uncertain, you still have an equal chance of falling either above or below the median. Therefore on an average, the outcome will be the same.'

24. A retailer uses two different formulas for predicting monthly sales. The first formula has an average miss of 700 records, and a standard deviation of 35 records. The second formula has an average miss of 300 records, and a standard deviation of 16. Which formula is relatively less accurate?

Self-practice Problems 4B

4.9 Find the average deviation from mean for the following distribution:

Quantity demanded (in units):	60	61	62	63	64	65	66	67	68
Frequency:	2	0	15	29	25	12	10	4	3

4.10 Find the average deviation from mean for the following distribution:

Dividend yield :	0–3	3–6	6–9	9–12	12–15	15–18	18–21
Number of companies:	2	7	10	12	9	6	4

4.11 Find the average deviation from median for the following distribution:

Sales (₹ '000) :	1–3	3–5	5–7	7–9	9–11	11–13	13–15	15–17
Number of shops :	6	53	85	56	21	26	4	4

4.12 In a survey of 48 engineering companies, following data was collected:

Level of profit (₹ in lakh) :	10	11	12	13	14
Number of companies :	3	12	18	12	3

Calculate the variance and standard deviation for the distribution.

4.13 A manufacturer of T-shirts approaches you with the following information

Length of Shoulder (in inches):	12.0	12.5	13.0	13.5	14	14.5	15	15.5	16
Frequency:	5	20	30	43	60	56	37	16	3

Calculate the standard deviation and advice the manufacturer as to the largest and the smallest shoulder size T-shirts he should make in order to meet the needs of his customers.

4.14 A charitable organization decided to give old-age pension to people over sixty years of age. The scales of pension were fixed as follows:

Age Group	*Pension/month* (₹)
60 – 65	200
65 – 70	250
70 – 75	300
75 – 80	350
80 – 85	400

The ages of 25 persons who secured the pension are as given below:

74	62	84	72	61	83	72	81	64
71	63	61	60	67	74	64	79	73
75	76	69	68	78	66	67		

Calculate the monthly average pension payable per person and the standard deviation.

4.15 Two automatic filling machines A and B are used to fill tea in 500 g cartons. A random sample of 100 cartons on each machine showed the following:

Tea Contents (in g)	*Machine A*	*Machine B*
485 – 490	12	10
490 – 495	18	15
495 – 500	20	24
500 – 505	22	20
505 – 510	24	18
510 – 515	4	13

Comment on the performance of the two machines on the basis of average filling and dispersion.

4.16 An analysis of production rejects resulted in the following observations

No. of Rejects per Operator	*No. of Operators*	*No. of Rejects per Operator*	*No. of Operators*
21 – 25	5	41 – 45	15
26 – 30	15	46 – 50	12
31 – 35	28	51 – 55	3
36 – 40	42		

Calculate the mean and standard deviation.

[*Delhi Univ., MBA, 2004*]

Manufacturer B : $\bar{x}_2 = 21.81$, $\sigma_2 = 7.074$ and C.V. = 32.44 per cent;

(a) Bags of manufacturer B have higher bursting pressure;

(b) Bags of manufacturer A have more uniform pressure;

(c) Bags of manufacturer A should be preferred by buyer as they have uniform pressure.

4.29 (a) CV(A) = 2.5 and CV(B) = 4.7. Variation in the distribution of daily wages per employee in factory B is more.

(b) Correct $\Sigma x = 100 \times 85 - 120 + 100 = 8{,}480$

Correct mean, $\bar{x} = 8480/100 = 84.8$

$$\begin{aligned}
\text{Since} \quad \sigma^2 &= (\Sigma x^2/\text{N}) - (\bar{x})^2\\
\text{or} \quad 16 &= (\Sigma x^2/100) - (85)^2\\
&= \Sigma x^2 - 7{,}22{,}500\\
\text{or} \quad \Sigma x^2 &= 7{,}24{,}100\\
\text{Correct} \quad \Sigma x^2 &= 7{,}24{,}100 - (120)^2 + (100)^2\\
&= 7{,}19{,}700\\
\text{Correct} \quad \sigma^2 &= (7{,}19{,}700/100) - (84.8)^2 = 5.96
\end{aligned}$$

4.30 CV(Mumbai) = 7.24 per cent ; CV (Kolkata) = 8.48 per cent. This shows more stability in Mumbai stock market.

4.31 CV(A) = 46.02 per cent ; CV(B) = 55 per cent. The purchase of petrol is relatively more variable at station B.

Formulae Used

1. Range, R

Value of highest observation – Value of lowest observation = H – L

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

2. Interquartile range = $Q_3 - Q_1$

$$\text{Quartile deviation, QD} = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3. Mean average deviation

For ungrouped data

(i) $\text{MAD} = \dfrac{\sum |x - \bar{x}|}{n}$, for sample

(ii) $\text{MAD} = \dfrac{\sum |x - \mu|}{\text{N}}$, for population

(iii) $\text{MAD} = \dfrac{\sum |x - \text{Me}|}{n}$, from median

For grouped data $\quad \text{MAD} = \dfrac{\sum f|x - \bar{x}|}{\sum f}$

4. $\text{Coefficient of MAD} = \dfrac{\text{MAD}}{\bar{x}\text{ or Me}} \times 100$

5. Variance

Ungrouped data

$$\sigma^2 = \frac{\sum (x-\bar{x})^2}{\text{N}} = \frac{\sum x^2}{\text{N}} - \left(\frac{\sum x}{\text{N}}\right)^2$$
$$= \frac{\sum d^2}{\text{N}} - \left(\frac{\sum d}{\text{N}}\right)^2$$

where $d = x - \text{A}$; A is any assumed A.M. value

$$\text{Grouped data, } \sigma^2 = \left[\frac{\sum f d^2}{\text{N}} - \left(\frac{\sum f d}{\text{N}}\right)^2\right] h$$

where $d = (m - \text{A})/h$; h is the class interval and m is the mid-value of class intervals.

6. Standard deviation

Ungrouped data, $\sigma = \sqrt{\sigma^2}$

$$\text{Grouped data, } \sigma = \sqrt{\frac{\sum f d^2}{\text{N}} - \left(\frac{\sum fd}{\text{N}}\right)^2} \times h$$

7. $\text{Coefficient of variation (CV)} = \dfrac{\sigma}{\bar{x}} \times 100$

Chapter Concepts Quiz

True or False

1. [T] [F] Range is a measure of variation which gives us information about scatter of values around a measure of central tendency.

2. [T] [F] When a distribution consists of different observations, s or σ are relatively large.

3. [T] [F] The interquartile range is based upon only two values in the data set.

4. [T] [F] Absolute measures of variation are used for comparing variability among observations in a data set.

5. [T] [F] The semi-interquartile range is inappropriate to use with skewed distributions.

6. [T] [F] Mean absolute deviation taken from median is least.

7. [T] [F] The standard deviation is measured in the same unit as the observations in the data set.

8. [T] [F] In a symmetrical distribution, semi-interquartile range is one-fourth of the range.

9. [T] [F] The coefficient of variation is a relative measure of dispersion.