
UNIT 3 MEASURES OF DISPERSION

Structure

- 3.0 Introduction
- 3.1 Objectives
- 3.2 Basic Measures of Dispersion
 - 3.2.1 Range
 - 3.2.2 Quartile Deviation
 - 3.2.3 Mean Deviation
- 3.3 Variance and Standard Deviation
 - 3.3.1 Variance
 - 3.3.2 Standard Deviation
 - 3.3.3 Root Mean Square Deviation
 - 3.3.4 Coefficient of Variation
- 3.4 Normal Data Sets
- 3.5 Concept of Skewness
- 3.6 Various Measures of Skewness
- 3.7 Concept of Kurtosis
- 3.8 Summary
- 3.9 Solutions / Answers
- 3.10 References

3.0 INTRODUCTION

Different measures of central tendency, discussed in Unit 2 of this block, give a value around which the data is concentrated. But it gives no idea about the nature of scatter or spread. For example, the observations 10, 30 and 50 have mean 30 while the observations 28, 30, 32 also have mean 30. Both the distributions are spread around 30. But it is observed that the variability among units is more in the first than in the second. In other words, there is greater variability or dispersion in the first set of observations in comparison to other. Measure of dispersion is calculated to get an idea about the variability in the data.

In Section 3.2, some basic measures of dispersion are described. Significance and properties of a good measure of dispersion are also explored. The basic idea about the range is explained in Section 3.2.1 whereas the measures of quartile deviation are introduced in Section 3.2.2. Mean deviation is described in Section 3.2.3. Methods of calculation of variance and standard deviation for grouped and ungrouped data are explained in Section 3.3 whereas in Section 3.3.3 the basis idea about root mean square deviation is provided. Coefficient of variation, which is a relative measure of variation is described in Section 3.3.4. Section 3.4 briefly defines the normal data distribution. Lack of symmetry of a frequency distribution is called skewness. If the distribution is not symmetric, the frequencies will not be normally distributed about the centre of the distribution. This unit also explains the concepts of skewness are described in Section 3.5 whereas the various measures of skewness are given with examples in Section 3.6. In Section 3.7, the concepts and a measure of kurtosis is discussed.

3.1 OBJECTIVES

After reading this unit, you would be able to

- conceptualise dispersion and its utility;
- explain the properties of a good measure of dispersion;
- explain methods of calculation of different types of measures of dispersion;
- solve the numerical problems related to the measures of dispersion.
- describe the properties of normal distribution;
- describe the concepts of skewness and its measures;
- explain the different measures of skewness;
- describe the concepts and a measure of kurtosis;
- explain how skewness and kurtosis describe the shape of a distribution.

3.2 BASIC MEASURES OF DISPERSION

According to Spiegel, the degree to which numerical data tend to spread about an average value is called the variation or dispersion of data. Actually, there are two basic kinds of measures of dispersion: (i) Absolute measures and (ii) Relative measures. The absolute measures of dispersion are used to measure the variability of a given data expressed in the same unit, while the relative measures are used to compare the variability of two or more sets of observations. Following are the different measures of dispersion:

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation and Variance

What is the Significance of Measures of Dispersion?

Measures of dispersion are needed for the following four basic purposes:

1. Measures of dispersion determine the reliability of an average value. In other words, measures of variation are pointed out as to how far an average is representative of the entire data. When variation is less, the average closely represents the individual values of the data and when variation is large; the average may not closely represent all the units and be quite unreliable.
2. Another purpose of measuring variation is to determine the nature and causes of variations in order to control the variation itself. For example, the variation in the quality of product in the process form of production can be checked by quality control department by identifying the reason for the variations in the quality of product. Thus, measurements of dispersion are helpful to control the causes of variation.
3. The measures of dispersion enable us to compare two or more series with regard to their variability. The relative measures of dispersion may also determine the uniformity or consistency. Smaller value of relative measure of dispersion implies greater uniformity or consistency in the data.

4. Measures of dispersion facilitate the use of other statistical methods. In other words, many powerful statistical tools in statistics such as correlation analysis, the testing of hypothesis, the analysis of variance, techniques of quality control, etc. are based on different measures of dispersion.

Properties of a Good Measure of Dispersion

A good measure of dispersion should possess the following properties:

1. It should be simple to understand;
2. It should be easy to compute;
3. It should be rigidly defined;
4. It should be based on each and every observation of data;
5. It should be amenable to further algebraic treatment;
6. It should have sampling stability; and
7. It should not be unduly affected by extreme observations.

Different measures of dispersion are discussed in following sub-sections.

3.2.1 Range

Range is the simplest measure of dispersion. It is defined as the difference between the maximum value of the variable and the minimum value of the variable in the distribution. Its merit lies in its simplicity. The demerit is that it is a crude measure because it is using only the maximum and the minimum observations of variable. However, it still finds applications in Order Statistics and Statistical Quality Control. It can be defined as

$$R = X_{\text{Max}} - X_{\text{Min}}$$

where, X_{Max} : Maximum value of variable and

X_{Min} : Minimum value of variable.

Example 1: Find the range of the distribution 6, 8, 2, 10, 15, 5, 1, 13.

Solution: For the given distribution, the maximum value of variable is 15 and the minimum value of variable is 1. Hence range = 15 - 1 = 14.

Note: Range is the simplest measure of dispersion as it can be obtained from the largest and the smallest observations. It is used in statistical quality control. However, since it uses only the maximum and the minimum values of a variable in the series and gives no importance to other observations, it is not very suitable for algebraic treatment. Range is affected by fluctuations in sampling. For example, if a single value lower than the minimum or higher than the maximum is added or if the maximum or minimum value is deleted range is seriously affected.

Range is the measure having unit of the variable but is not a pure number. That's why sometimes coefficient of range is calculated by

$$\text{Coefficient of Range} = \frac{X_{\text{Max}} - X_{\text{Min}}}{X_{\text{Max}} + X_{\text{Min}}}$$

3.2.2 Quartile Deviation

You have already studied in about Q_1 and Q_3 , the first quartile and the third quartile respectively in the Unit 2. ($Q_3 - Q_1$) gives the interquartile range. The

semi-interquartile range which is also known as Quartile Deviation (QD) is given by

$$\text{Quartile Deviation (QD)} = (Q_3 - Q_1) / 2$$

Relative measure of Q.D. known as Coefficient of Q.D. and is defined as

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

Example 2: For the following data, find the quartile deviation:

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	03	05	07	09	04

Solution: We have $N/4 = 28/4 = 7$ and 7th observation falls in the class 10-20. This is the first quartile class.

Similarly, $3N/4 = 21$ and 21st observation falls in the interval 30-40. This is the third quartile class.

Class Interval	Frequency	Cumulative Frequency
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28

Using the formulae of first quartile and third quartile we found

$$Q_1 = 10 + \frac{(7-3)}{5} \times 10 = 18$$

$$Q_3 = 30 + \frac{(21-15)}{9} \times 10 = 36.67$$

Hence

$$\text{Quartile Deviation} = (36.67 - 18) / 2 = 9.335$$

3.2.3 Mean Deviation

Mean deviation is defined as average of the sum of the absolute values of deviation from any arbitrary value viz. mean, median, mode, etc. It is often suggested to calculate it from the median because it gives least value when measured from the median.

The deviation of an observation x_i from the assumed mean A is defined as $(x_i - A)$.

Therefore, the mean deviation can be defined as

$$MD = \frac{1}{n} \sum_{i=1}^n |x_i - A|$$

The quantity $\sum |x_i - A|$ is minimum when A is median.

We accordingly define mean deviation from mean as

$$MD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

and from the median as

$$MD = \frac{\sum_{i=1}^n |x_i - \text{median}|}{n}$$

For frequency distribution, the formula will be

$$MD = \frac{\sum_{i=1}^k f_i |x_i - \bar{x}|}{\sum_{i=1}^k f_i}$$

$$MD = \frac{\sum_{i=1}^k f_i |x_i - \text{median}|}{\sum_{i=1}^k f_i}$$

where, all symbols have usual meanings.

Example 3: Find mean deviation for the given data

1, 2, 3, 4, 5, 6, 7

Solution: First of all we find Mean

$$\bar{x} = \frac{1+2+3+4+5+6+7}{7} = \frac{28}{7} = 4$$

Then, we will find $|x_i - \bar{x}|$: 3, 2, 1, 0, 1, 2, 3

So, $\sum |x_i - \bar{x}| = 12$

Therefore,

$$MD = \frac{12}{7} = 1.71$$

Example 4: Find mean deviation from mean for the following data:

x	1	2	3	4	5	6	7
f	3	5	8	12	10	7	5

Solution: First of all we have to calculate the mean from the given data

x	f	fx	$ x - \bar{x} $	$f x - \bar{x} $
1	3	3	3.24	9.72
2	5	10	2.24	11.20
3	8	24	1.24	9.92
4	12	48	0.24	2.88
5	10	50	0.76	7.60
6	7	42	1.76	12.32
7	5	35	2.76	13.80
	50	212	12.24	67.44

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{212}{50} = 4.24$$

$$MD = \frac{\sum_{i=1}^k f_i |x_i - \bar{x}|}{\sum_{i=1}^k f_i} = \frac{67.44}{50} = 1.348$$

Example 5: Consider the data given in Example 2 and find the Mean deviation from median.

Solution:

Class	Mid Value (x)	Frequency (f)	C.F.	$f_i x_i - \text{median} $
0-10	5	3	3	70.71
10-20	15	5	8	67.85
20-30	25	7	15	24.99
30-40	35	9	24	57.87
40-50	45	4	28	65.72
		$\Sigma f = 28$		$\Sigma f_i x_i - \text{median} = 287.14$

We have $N/2 = 28/2 = 14$

The 14th observation falls in the class 20-30. This is therefore the median class.

Using the formula of median, $\text{Median} = 20 + \frac{14-8}{7} \times 10 = 28.57$

$$\text{Mean Deviation from Median} = \frac{\sum_{i=1}^k f_i |x_i - \text{median}|}{\sum_{i=1}^k f_i} = 10.255$$

Note: Mean deviation utilises all the observations and is easy to understand and calculate. It is not much affected by extreme values. However, while calculating mean deviation, the negative deviations are made positive, thus, mean deviation is not amenable to algebraic treatment.

Check Your Progress 1:

1) Calculate the range for the following data:

60, 65, 70, 12, 52, 40, 48

.....

.....

2) Calculate range for the following frequency distribution:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	2	6	12	7	3

.....

.....

3) Calculate the Quartile Deviation for the following data:

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	6	8	12	24	36	32	24	8

.....

.....

.....

.....

4) Following are the marks of 7 students in Statistics:

16, 24, 13, 18, 15, 10, 23

Find the mean deviation from mean.

.....

.....

.....

.....

5) Find mean deviation for the following distribution:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

.....

.....

.....

.....

3.3 VARIANCE AND STANDARD DEVIATION

In this section, we discuss one of the most used measures of dispersion and certain related measures. We will first discuss the variance, followed by the standard deviation and the root mean square variation.

3.3.1 Variance

Variance is the average of the square of deviations of the values taken from mean. Taking a square of the deviation is a better technique to get rid of negative deviations.

Variance is defined as

$$\text{Var}(x) = \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

and for a frequency distribution, the formula is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^k f_i (x_i - \bar{x})^2$$

where, all symbols have their usual meanings.

It should be noted that sum of squares of deviations is least when deviations are measured from the mean. This means $\sum (x_i - A)^2$ is least when $A = \text{Mean}$.

Example 6: Calculate the variance for the data given in Example 3.

Solution:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
1	-3	9
2	-2	4
3	-1	1
4	0	0
5	1	1
6	2	4
7	3	9

We have
$$\sum_{i=1}^n (x_i - \bar{x})^2 = 28$$

Therefore,
$$\text{Var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{7} \times 28 = 4$$

Example 7: For the data given in Example 2, compute the variance.

Solution: We have the following data:

Class	Mid Value	Frequency (f)	$f(x - \bar{x})^2$
0-10	5	3	1470.924
10-20	15	5	737.250
20-30	25	7	32.1441
30-40	35	9	555.606
40-50	45	4	1275.504
		$\sum f = 28$	$\sum f(x_i - \bar{x})^2$ =4071.428

$$\begin{aligned} \text{Variance} &= \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{\sum_{i=1}^k f_i} \\ &= 4071.429 / 28 = 145.408 \end{aligned}$$

Note: A Variance is a rigidly defined measure that utilises all the observations. It is amenable to algebraic treatment and get rid of negative deviations by squaring. It is the most popular measure of dispersion. However, in cases where mean is not a suitable average, like when open ended classes are present, variance may not be the coveted measure of dispersion. In such cases quartile deviation may be used. Further, its unit is the square of the unit of the variable, due to which it is difficult to judge the magnitude of dispersion compared to standard deviation.

Variance of Combined Series

If there are two or more populations and the information about the means and variances of those populations are available then we can obtain the combined variance of several populations. If $n_1, n_2, \dots, n_k$ are the sizes, $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ are the means and $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$ are the variances of k populations, then the combined variance is given by

$$\sigma^2 = \frac{1}{(n_1 + n_2 + \dots + n_k)} [n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + \dots + n_k(\sigma_k^2 + d_k^2)]$$

where, $d_i = \bar{x}_i - \bar{x}$ and

$$\bar{x} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{\sum_{i=1}^k n_i} \text{ is the mean of the combined series.}$$

Example 8: Suppose a series of 100 observations has mean 50 and variance 20. Another series of 200 observations has mean 80 and variance 40. What is the combined variance of the given series?

Solution: First we find the mean of the combined series

$$\begin{aligned} \bar{x} &= \frac{(100 \times 50 + 200 \times 80)}{(100 + 200)} \\ &= \frac{21000}{300} = 70 \end{aligned}$$

Therefore, $d_1 = 50 - 70 = -20$ and $d_2 = 80 - 70 = 10$

Variance of the combined series

$$\begin{aligned} &= \frac{(100 \times (20 + 400) + 200 \times (40 + 100))}{(100 + 200)} \\ &\Rightarrow \frac{(42000 + 28000)}{(300)} = \frac{70000}{300} = 233.33 \end{aligned}$$

3.3.2 Standard Deviation

Standard deviation (SD) is defined as the positive square root of variance. The formula is

$$SD = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

and for a frequency distribution the formula is

$$SD = \sqrt{\frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{\sum_{i=1}^k f_i}}$$

where, all symbols have usual meanings. SD, MD and variance cannot be negative.

Example 9: Find the SD for the data in Example 2.

Solution: In the Example 7, we have already found the

$$\text{Variance} = 145.408$$

So

$$\text{SD} = +\sqrt{(145.408)} = 12.04$$

Note: Standard deviation is a rigidly defined measure that utilises all the observations. It is amenable to algebraic treatment and get rid of negative deviations by squaring. It is the most popular measure of dispersion. However, in cases where mean is not a suitable average, like when open ended classes are present, variance may not be the coveted measure of dispersion. In such cases quartile deviation may be used.

Remark:

SD is algebraically more amenable than MD. MD straightaway neglects the sign of observations which is also a useful information. In SD, the signs are considered but they get automatically eliminated due to squaring. However, MD will be useful in situations where Median is more suitable than mean as an average.

3.3.3 Root Mean Square Deviation

As we have discussed in the last sub-section, standard deviation is the positive square root of the average of the squares of deviations taken from the mean. If we take the deviations from assumed mean then it is called Root Mean Square Deviation and it is defined as

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^n (x_i - A)^2}{n}}$$

where, A is the assumed mean.

For a frequency distribution, the formula is

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^k f_i (x_i - A)^2}{\sum_{i=1}^k f_i}}$$

When assumed mean is equal to the actual mean $\bar{x}$ i.e. $A = \bar{x}$ root mean square deviation will be equal to the standard deviation.

3.3.4 Coefficient Of Variation

Coefficient of Variation (CV) is defined as

$$\text{CV} = \frac{\sigma}{\bar{x}} \times 100$$

It is a relative measure of variability. If we are comparing the two data series, the data series having smaller CV will be more consistent. One should be careful in making interpretation with CV. For example, the series 10, 10, 10 has SD zero and hence CV is also zero. The series 50, 50 and 50 also has SD zero and hence

CV is zero. But the second series has higher mean. So we shall regard the second series as more consistent than the first.

Example 10: Suppose batsman A has mean 50 with SD 10. Batsman B has mean 30 with SD 3. What do you infer about their performance?

Solution: A has higher mean than B. This means A is a better run maker.

However, B has lower CV ($3/30 = 0.1$) than A ($10/50 = 0.2$) and is consequently more consistent.

Check Your Progress 2

- 1) For a group containing 100 observations the arithmetic mean and standard deviation are 16 and $\sqrt{21}$ respectively. For 50 observations, the mean and standard deviation are 20 and 2 respectively. Calculate mean and standard deviation of other half.

.....

- 2) Find the standard deviation for the following numbers:

10 27 40 60 33 30 10

.....

- 3) Calculate standard deviation for the following data:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

.....

- 4) If $n = 10$, $\bar{x} = 4$, $\sum x^2 = 200$, find the coefficient of variation.

.....

- 5) For a distribution, the coefficient of variation is 70.6% and the value of arithmetic mean is 16. Find the value of standard deviation.

.....

3.4 NORMAL DATA SETS

The frequency distribution of the normal data is of the shape of a 'Bell' and is symmetrical on both sides. This bell-shaped curve, also known as the Normal Curve, has identical values of all three measures of central tendency (Mean, Median and Mode). This frequency distribution is known as the Normal Distribution. Many data variables, like height, weight, age, blood sugar level, etc., are found to be normally distributed in nature.

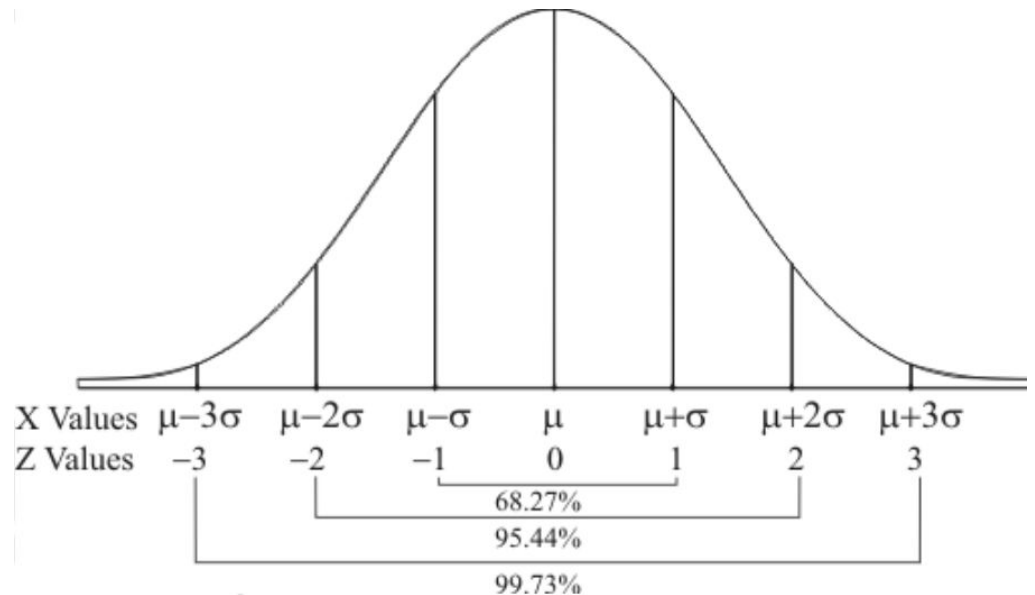


Fig. 3.1: Normal Distribution

The following are the characteristics of a normal distribution:

1. The Normal Curve is Symmetrical: The normal probability curve is symmetrical around its vertical axis, called the ordinate. The symmetry about the ordinate at the central point of the curve implies that the left and right halves of the middle central point are mirror images (Refer to Fig. 3.1).
2. The Normal Curve is Unimodal: Since there is only one maximum point in the curve, the normal probability curve is unimodal, i.e. it has only one mode.
3. The Maximum Ordinate occurs at the Centre: The maximum height of the ordinate always occurs at the central point of the curve, that is, the mid-point.
4. The Normal Curve is Asymptotic to the X Axis: The normal probability curve approaches the horizontal axis asymptotically; i.e. the curve continues to decrease in height on both ends away from the middle point (the maximum ordinate point), but it never touches the horizontal axis. Therefore, its ends extend from minus infinity ($-\infty$) to plus infinity ($+\infty$).
5. The Height of the Curve declines Symmetrically: In the normal probability curve, the height declines symmetrically in either direction from the maximum point.
6. The Points of Influx occur at point ± 1 Standard Deviation ($\pm 1 \sigma$): The normal curve changes its direction from convex to concave at a point recognised as point of influx.
7. The Total Percentage of Area of the Normal Curve within Two Points of inflexion is fixed: Approximately 68.26 of % area of the curve lies within the limits of ± 1 standard deviation ($\pm 1 \sigma$) unit from the mean.
8. The Total Area under Normal Curve may also be considered 100 Per cent Probability: The total area under the normal curve may be considered to approach 100 per cent probability, interpreted in terms of standard deviations.
9. The Normal Curve is Bilateral: The 50% area of the curve lies to the left side of the maximum central ordinate, and 50% of the area lies to the right side. Hence, the curve is bilateral.

Any deviation from the normal distribution of data is measured by the skewness and kurtosis, which are defined in the subsequent sections.

3.5 CONCEPT OF SKEWNESS

Skewness means lack of symmetry. In mathematics, a figure is called symmetric if there exists a point in it through which if a perpendicular is drawn on the X-axis, it divides the figure into two congruent parts i.e. identical in all respect or one part can be superimposed on the other i.e. mirror images of each other. In Statistics, a distribution is called symmetric if mean, median and mode coincide. Otherwise, the distribution becomes asymmetric. If the right tail is longer, we get a positively skewed distribution for which $\text{mean} > \text{median} > \text{mode}$ while if the left tail is longer, we get a negatively skewed distribution for which $\text{mean} < \text{median} < \text{mode}$.

The example of the Symmetrical curve, Positive skewed curve and Negative skewed curve are given as follows:

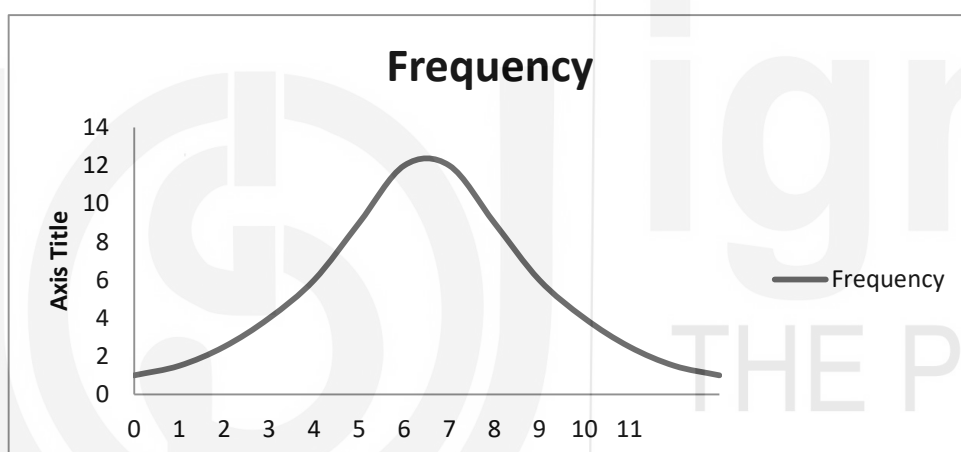


Fig. 3.2: Symmetrical Curve

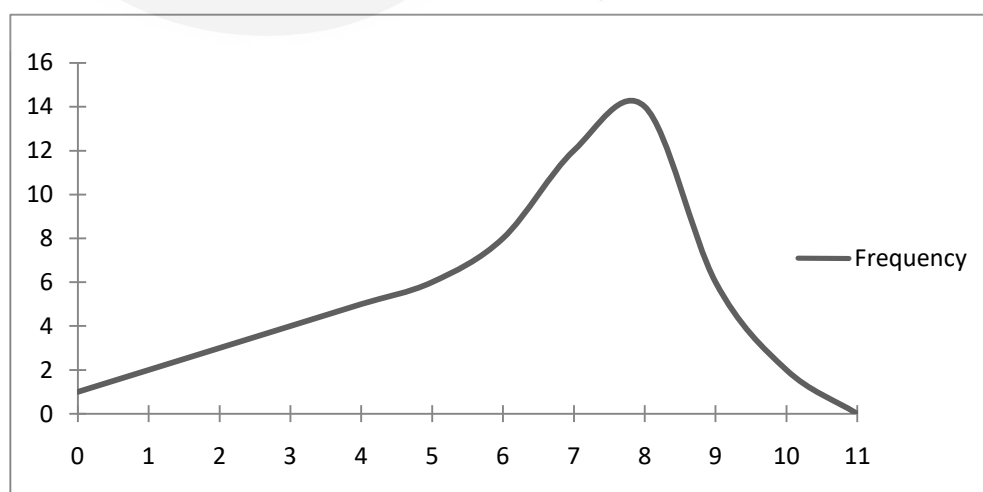


Fig. 3.3: Negative Skewed Curve

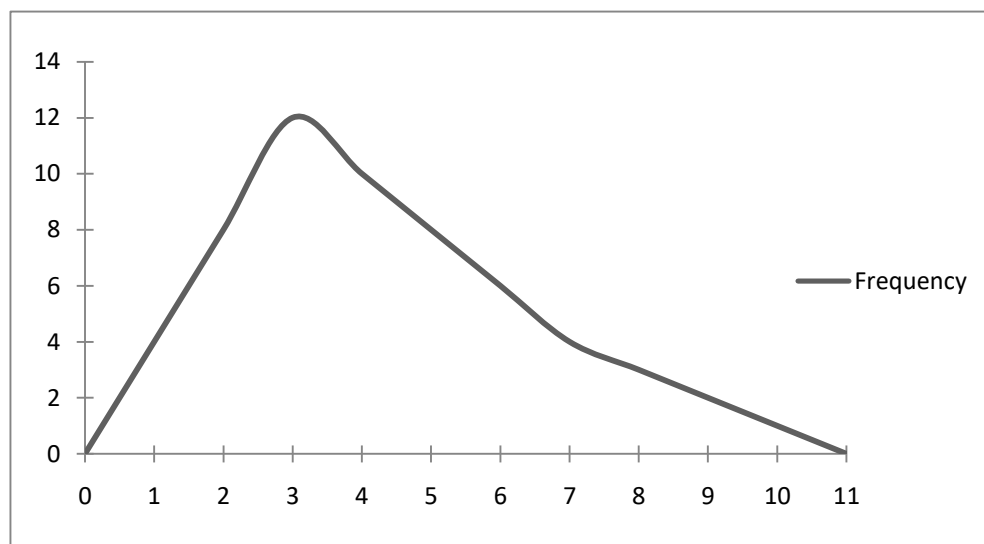


Fig. 3.4: Positive Skewed Curve

Difference between Variance and Skewness

The following two points of difference between variance and skewness should be carefully noted.

1. Variance tells us about the amount of variability while skewness gives the direction of variability.
2. In business and economic series, measures of variation have greater practical application than measures of skewness. However, in medical and life science field measures of skewness have greater practical applications than the variance.

3.6 VARIOUS MEASURES OF SKEWNESS

Measures of skewness help us to know to what degree and in which direction (positive or negative) the frequency distribution has a departure from symmetry. Although positive or negative skewness can be detected graphically depending on whether the right tail or the left tail is longer but, we don't get idea of the magnitude. Besides, borderline cases between symmetry and asymmetry may be difficult to detect graphically. Hence some statistical measures are required to find the magnitude of lack of symmetry. A good measure of skewness should possess three criteria:

1. It should be a unit free number so that the shapes of different distributions, so far as symmetry is concerned, can be compared even if the unit of the underlying variables are different;
2. If the distribution is symmetric, the value of the measure should be zero. Similarly, the measure should give positive or negative values according as the distribution has positive or negative skewness respectively; and
3. As we move from extreme negative skewness to extreme positive skewness, the value of the measure should vary accordingly.

Measures of skewness can be both absolute as well as relative. Since in a symmetrical distribution mean, median and mode are identical more the mean moves away from the mode, the larger the asymmetry or skewness. An absolute measure of skewness cannot be used for purposes of comparison because of the

same amount of skewness has different meanings in distribution with small variation and in distribution with large variation.

Absolute Measures of Skewness

Following are the absolute measures of skewness:

1. Skewness (S_k) = Mean – Median
2. Skewness (S_k) = Mean – Mode
3. Skewness (S_k) = ($Q_3 - Q_2$) - ($Q_2 - Q_1$)

For comparing to series, we do not calculate these absolute measures we calculate the relative measures which are called coefficient of skewness. Coefficient of skewness are pure numbers independent of units of measurements.

Relative Measures of Skewness

In order to make valid comparison between the skewness of two or more distributions we have to eliminate the distributing influence of variation. Such elimination can be done by dividing the absolute skewness by standard deviation. Skewness can be measured using the β and γ coefficients of Skewness, which were defined by Karl Pearson based upon the second and third central moments. Discussion on these is beyond the scope of this Unit and can be studied from the references.

In this section we discuss few simpler ways of defining Skewness.

1. Karl Pearson's Coefficient of Skewness

This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

The value of this coefficient would be zero in a symmetrical distribution. If mean is greater than mode, coefficient of skewness would be positive otherwise negative. The value of the Karl Pearson's coefficient of skewness usually lies between ± 1 for moderately skewed distribution. If mode is not well defined, we use the formula

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

By using the relationship

$$\text{Mode} = (3 \text{ Median} - 2 \text{ Mean})$$

Here, $-3 \leq S_k \leq 3$. In practice it is rarely obtained.

2. Bowleys's Coefficient of Skewness

This method is based on quartiles. The formula for calculating coefficient of skewness is given by

$$S_k = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

$$= \frac{(Q_3 - 2Q_2 + Q_1)}{(Q_3 - Q_1)}$$

The value of S_k would be zero if it is a symmetrical distribution. If the value is greater than zero, it is positively skewed and if the value is less than zero it is negatively skewed distribution. It will take value between +1 and -1.

3. Kelly's Coefficient of Skewness

The coefficient of skewness proposed by Kelly is based on percentiles and deciles. The formula for calculating the coefficient of skewness is given by

Based on Percentiles

$$\begin{aligned} S_k &= \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{10})} \\ &= \frac{(P_{90} - 2P_{50} + P_{10})}{(P_{90} - P_{10})} \end{aligned}$$

where, P_{90} , P_{50} and P_{10} are 90th, 50th and 10th Percentiles.

Based on Deciles

$$S_k = \frac{(D_9 - 2D_5 + D_1)}{D_9 - D_1}$$

where, D_9 , D_5 and D_1 are 9th, 5th and 1st Decile.

Example 11: For a distribution, Karl Pearson's coefficient of skewness is 0.64, standard deviation is 13, and mean is 59.2. Find the mode and median.

Solution: We have given

$$S_k = 0.64, \sigma = 13 \text{ and Mean} = 59.2$$

Therefore by using formulae

$$\begin{aligned} S_k &= \frac{\text{Mean} - \text{Mode}}{\sigma} \\ 0.64 &= \frac{59.2 - \text{Mode}}{13} \end{aligned}$$

$$\text{Mode} = 59.20 - 8.32 = 50.88$$

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

$$50.88 = 3 \text{ Median} - 2 (59.2)$$

$$\text{Median} = \frac{50.88 + 118.4}{3} = \frac{169.28}{3} = 56.42$$

Remarks about Skewness

1. If the value of mean, median and mode are same in any distribution, then the skewness does not exist in that distribution. Larger the difference in these values, larger the skewness;
2. If sum of the frequencies are equal on the both sides of mode then skewness does not exist;
3. If the distance of first quartile and third quartile are same from the median then a skewness does not exist. Similarly if deciles (first and ninth) and

percentiles (first and ninety nine) are at equal distance from the median. Then there is no asymmetry;

4. If the sums of positive and negative deviations obtained from mean, median or mode are equal then there is no asymmetry; and
5. If a graph of a data become a normal curve and when it is folded at middle and one part overlap fully on the other one then there is no asymmetry.

3.7 CONCEPT OF KURTOSIS

If we have the knowledge of the measures of central tendency, dispersion and skewness, even then we cannot get a complete idea of a distribution. In addition to these measures, we need to know another measure to get the complete idea about the shape of the distribution which can be studied with the help of Kurtosis. Prof. Karl Pearson has called it the “Convexity of a Curve”. Kurtosis gives a measure of flatness of distribution.

The degree of kurtosis of a distribution is measured relative to that of a normal curve. The curves with greater peakedness than the normal curve are called “**Leptokurtic**”. The curves which are more flat than the normal curve are called “**Platykurtic**”. The normal curve is called “**Mesokurtic**.” The Fig. 3.5 describes the three different curves mentioned above:

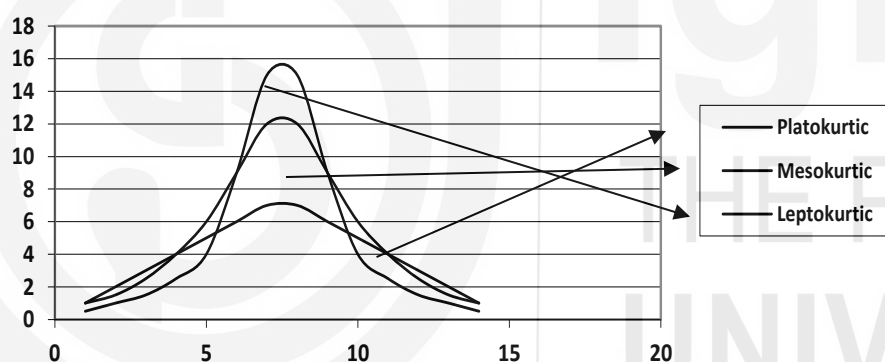


Fig. 3.5: Platykurtic Curve, Mesokurtic Curve and Leptokurtic Curve

Here we discuss one simple measure of Kurtosis. You can refer to other measures of kurtosis from references.

Kelly's Measure of Kurtosis

Kelly has given a measure of kurtosis based on percentiles. The formula is given by

$$\beta_2 = \frac{P_{75} - P_{25}}{P_{90} - P_{10}}$$

where, P_{75} , P_{25} , P_{90} , and P_{10} are 75th, 25th, 90th and 10th percentiles of dispersion respectively.

If $\beta_2 > 0.26315$, then the distribution is platykurtic.

If $\beta_2 < 0.26315$, then the distribution is leptokurtic.

Check Your Progress 3

- 1) Karl Pearson's coefficient of skewness is 1.28, its mean is 164 and mode 100, find the standard deviation.
-
-

- 2) For a frequency distribution the Bowley's coefficient of skewness is 1.2. If the sum of the 1st and 3rd quarterlies is 200 and median is 76, find the value of third quartile.
-
-
-

- 3) The following are the marks of 150 students in an examination. Calculate Karl Pearson's coefficient of skewness.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Students	10	40	20	0	10	40	16	14

.....

.....

.....

.....

3.8 SUMMARY

In this unit, we have discussed:

1. A point to remember - While the mean value gives the value around which a distribution is scattered, the measure of dispersion tells how it is scattered.
2. This unit discusses the measures of dispersion, its use and measures of dispersion
3. The properties of a good measure of dispersion are also explained.
4. The different types of measures of dispersion were also discussed.
5. A suitable measure of central tendency and dispersion is calculated (say, arithmetic mean and SD), we get a good idea of the distribution even if the distribution is large.
6. This unit also defines the normal data and properties of normal distribution
7. It defines the significance of skewness and different kinds of measures of skewness.
8. The unit also explains the concept of kurtosis and different measures of kurtosis

3.9 SOLUTIONS /ANSWERS**Check Your Progress 1**

- 1) Maximum value (X_{Max}) = 70
 Minimum value (X_{Min}) = 12
 Therefore, $R = X_{\text{Max}} - X_{\text{Min}}$
 $= 70 - 12 = 58$

- 2) In the given frequency distribution, we have

Lower limit of first class = 0

Upper limit of last class = 50

Therefore,

$$\text{Range} = 50 - 0 = 50$$

- 3) First we construct the following cumulative frequency distribution:

Class	f	Cumulative frequencies
0-5	6	6
5-10	8	14
10-15	12	26
15-20	24	50
20-25	36	86
25-30	32	118
30-35	24	142
35-40	8	150
	N=150	

$$Q_1 = L + \frac{\left(\frac{N}{4} - C\right)}{f} \times h$$

15-20 is the Q_1 class

$$= 15 + \frac{\left(\frac{150}{4} - 26\right)}{24} \times 5$$

$$= 15 + \frac{37.5 - 26}{24} \times 5 = 15 + \frac{57.5}{24}$$

$$= 17.40$$

$$Q_3 = L + \frac{\left(\frac{3N}{4} - C\right)}{f} \times h$$

25-30 is the Q_3 class

$$= 25 + \frac{\left(\frac{3 \times 150}{4} - 86\right)}{32} \times 5$$

$$= 25 + \frac{112.5 - 86}{24} \times 5 = 25 + \frac{132.5}{24}$$

$$= 30.52$$

$$\text{Therefore, } QD = \frac{Q_3 - Q_1}{2} = \frac{30.52 - 17.40}{2} = \frac{13.12}{2} = 6.56$$

- 4) We have the following distribution:

x	(x-17)	$ x - \bar{x} $
16	-1	1
24	+7	7
13	-4	4

18	+1	1
15	-2	2
10	-7	7
23	+6	6
$\sum x = 119$		$\sum x - \bar{x} = 28$

Then
$$\bar{x} = \frac{\sum x}{n} = \frac{119}{7} = 17$$

Therefore,
$$MD = \frac{\sum |x - \bar{x}|}{n} = \frac{28}{7} = 4$$

5) First we construct the following frequency distribution:

Class	x	f	xf	(x - $\bar{x}$)	x - $\bar{x}$	f x - $\bar{x}$
0-10	5	5	25	-22	22	110
10-20	15	8	120	-12	12	96
20-30	25	15	375	-2	2	30
30-40	35	16	560	8	8	128
40-50	45	6	270	18	18	108
						$\sum f x - \bar{x} = 472$

Then
$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{1350}{50} = 27$$

and

$$MD = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{472}{50} = 9.44$$

Check Your Progress 2

1) We have given

$$n = 100, \quad \bar{x} = 16, \quad \sigma = \sqrt{21}$$

$$n_1 = 50, \quad \bar{x}_1 = 20, \quad \sigma_1 = 2, \quad \text{and} \quad n_2 = 50$$

we have to find

$$\bar{x}_2 = ?, \quad \sigma_2 = ?$$

$$\text{Now, } \bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{50 \times 20 + 50 \bar{x}_2}{100}$$

$$16 = \frac{50 \times 20 + 50 \bar{x}_2}{100}$$

$$50 \bar{x}_2 = 16 \times 100 - 50 \times 20$$

$$50 \bar{x}_2 = 1600 - 1000 = 600$$

$$\Rightarrow \bar{x}_2 = 12$$

$$d_1 = \bar{x}_1 - \bar{x} = 20 - 16 = 4 \Rightarrow d_1^2 = 16$$

$$d_2 = \bar{x}_2 - \bar{x} = 12 - 16 = -4 \Rightarrow d_2^2 = 16$$

$$\sigma^2 = \frac{1}{n_1 + n_2} [n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$$

$$(\sqrt{21})^2 = \frac{[50(4+16) + 50(\sigma_2^2 + 16)]}{100}$$

$$21 \times 100 - 20 \times 50 = 50\sigma_2^2 + 50 \times 16$$

$$50\sigma_2^2 = 2100 - 1000 - 800$$

$$50\sigma_2^2 = 2100 - 1000 - 800$$

$$50\sigma_2^2 = 300$$

$$\Rightarrow \sigma_2^2 = 6 \Rightarrow \sigma_2 = \sqrt{6}$$

2) First we calculate the following distribution:

Then

x	x ²
10	100
27	729
40	1600
60	3600
33	1089
30	900
10	100
$\sum x = 210$	$\sum x^2 = 8118$

$$\bar{x} = \frac{\sum x}{n} = \frac{210}{7} = 30$$

and therefore,

$$\sigma = \sqrt{\frac{1}{n} \sum x^2 - (\bar{x})^2} = \sqrt{\frac{8118}{7} - 900}$$

$$= \sqrt{1159.7 - 900} = \sqrt{259.7} = 16.115$$

3) Let us take A = 25 and calculate the following frequency distribution:

Class	x	f	d = x-A	f d	f d ²
0-10	5	5	-20	-100	2000
10-20	15	8	-10	-80	800
20-30	25	15	0	0	0
30-40	35	16	10	160	1600
40-50	45	6	20	120	2400
		N = 50		$\sum fd = 100$	$\sum fd^2 = 6800$

Therefore,

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^k f_i d_i^2 - \left(\frac{\sum_{i=1}^k f_i d_i}{N} \right)^2} = \sqrt{\frac{1}{50} (6800) - \left(\frac{100}{50} \right)^2}$$

$$= \sqrt{136 - 4} = \sqrt{132} = 11.49$$

4) We have

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

$$= \frac{1}{n} \times 200 - (4)^2 = 20 - 16 = 4$$

$$\Rightarrow \sigma^2 = 4 \Rightarrow \sigma = 2$$

Therefore, coefficient of variation

$$C.V(X) = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{2}{4} \times 100 = 50\%$$

5) We have the formulae for coefficient of variation

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

By putting the given values, we have

$$\sigma = \frac{CV \times \bar{x}}{100} = \frac{70.6 \times 16}{100}$$

$$= 11.296$$

Check Your Progress 3

1) Using the formulae, we have

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$1.28 = \frac{164 - 100}{\sigma}$$

$$\sigma = \frac{64}{1.28} = 50$$

2) We have given $S_k = 1.2$

$$Q_1 + Q_3 = 200$$

$$Q_2 = 76$$

$$\text{then } S_k = \frac{(Q_3 + Q_1 - 2Q_2)}{(Q_3 - Q_1)}$$

$$1.2 = \frac{(200 - 2 \times 76)}{(Q_3 - Q_1)}$$

$$Q_3 - Q_1 = \frac{48}{1.2} = 40$$

$$Q_3 - Q_1 = 40$$

... (1)

and it is given $Q_1 + Q_3 = 200$

$$Q_1 = 200 - Q_3$$

Therefore, from equation (1)

$$Q_3 - (200 - Q_3) = 40$$

$$2Q_3 = 240$$

$$Q_3 = 120$$

- 3) Let us calculate the mean and median from the given distribution because mode is not well defined.

Class	f	x	CF	$d' = \frac{x-35}{10}$	fd'	fd' ²
0-10	10	5	10	-3	-30	90
10-20	40	15	50	-2	-80	160
20-30	20	25	70	-1	-20	20
30-40	0	35	70	0	0	0
40-50	10	45	80	+1	10	10
50-60	40	55	120	+2	80	160
60-70	16	65	136	+3	48	144
70-80	14	75	150	+4	56	244
					$\sum fd' = 64$	$\sum fd'^2 = 828$

$$\begin{aligned} \text{Median} &= L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h \\ &= 40 + \frac{75 - 70}{10} \times 10 = 45 \end{aligned}$$

$$\begin{aligned} \text{Mean } (\bar{x}) &= A + \frac{\sum fd'}{N} \times h \\ &= 35 + \frac{64}{150} \times 10 = 39.27 \end{aligned}$$

$$\begin{aligned} \text{Standard Deviation } (\sigma) &= h \times \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \\ &= 10 \times \sqrt{\frac{828}{150} - \left(\frac{64}{150}\right)^2} \\ &= 10 \times \sqrt{5.33} = 23.1 \end{aligned}$$

Therefore, coefficient of skewness:

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

$$= \frac{3(39.27 - 45)}{23.1} = -0.744$$

3.10 REFERENCES

IGNOU Course Material

1. Post-Graduate Diploma In Applied Statistics (PGDAST), MST 002: Descriptive Statistics, Block 1: Analysis of Quantitative Data, UNIT 2: Measures of Dispersion (for Sections 3.2, 3.3 and related portions in Sections 3.0, 3.1, 3.8, 3.9 and Check Your Progress questions)
2. PGDAST, MST 002: Descriptive Statistics, Block 1: Analysis of Quantitative Data, UNIT 3: Skewness and Kurtosis (for Sections 3.5 to 3.7 and related portions in Sections 3.0, 3.1, 3.8, 3.9 and Check Your Progress questions)
3. Master of Arts in Psychology (MAPC), MPC-006: Statistics in Psychology; Block 3: Normal Distribution, Unit 1: Characteristics of Normal Distribution (for Section 3.4 and related portions in Sections 3.0, 3.1, 3.8, 3.9 and Check Your Progress questions)

Reference Book

4. “Introduction to Probability and Statistics for Engineers and Scientists”, 3rd Edition, Sheldon M. Ross, Elsevier Academic Press, 2004