CHAPTER 8

Statistical Methods

Chapter Outline

8.1 Introduction
8.2 Data Analysis
8.3 Classification of Data
8.4 Frequency Distribution
8.5 Graphical Representation
8.6 Measures of Central Tendency
8.7 Arithmetic Mean
8.8 Median
8.9 Mode
8.10 Standard Deviation
8.11 Moments
8.12 Random Variables
8.13 Discrete Probability Distribution
8.14 Discrete Distribution Function
8.15 Measures of Central Tendency for a Discrete Probability Distribution

8.1 INTRODUCTION

Statistics is the science which deals with the collection, presentation, analysis, and interpretation of numerical data. Statistics should possess the following characteristics:

(i) Statistics are aggregates of facts.
(ii) Statistics are affected by a large number of causes.
(iii) Statistics are always numerically expressed.
(iv) Statistics should be enumerated or estimated.
(v) Statistics should be collected in a systematic manner.
(vi) Statistics should be collected for a pre-determined purpose.
(vii) Statistics should be placed in relation to each other.

The use of statistical methods help in presenting a complex mass of data in a simplified form so as to facilitate the process of comparison of characteristics in two or more situations. Statistics also provide important techniques for the study of relationship between two or more characteristics (or variables) in forecasting, testing of hypothesis, quality control, decision making, etc.

8.2 DATA ANALYSIS

The collection and analysis of data constitute the main stages of execution of any statistical investigation. The procedure for collection of data depends upon various considerations such as objective, scope, nature of investigation, etc. Data may be collected for each and every unit of the whole lot (population), which will ensure greater accuracy. Data may also be collected for a sample of population and conclusions that can be drawn on the basis of this sample are taken to hold for the population.

8.3 CLASSIFICATION OF DATA

The collected data are a complex and unorganized mass of figures which is very difficult to analyze and interpret. Therefore, it becomes necessary to organize the data so that it is easier to grasp its broad features. In order to analyze the data, it is essential that the data are arranged in a definite form. This task is accomplished by the process of classification. The main objectives of any classification are

(i) To present the data in a condensed form.
(ii) To bring out the relationship between variables.
(iii) To prepare data for tabulation and analysis.
(iv) To highlight the effect of one variable by eliminating the effect of others.

Consider the raw data relating to marks obtained in mathematics by a group of 60 students:

38, 11, 40, 0, 26, 15, 5, 40, 31, 12, 35, 0, 7, 20, 5, 28, 8, 21, 7, 28, 48, 45, 42, 17, 2, 38, 41, 18, 16, 16, 0, 19, 10, 7, 5, 1, 17, 22, 35, 44, 28, 46, 9, 16, 29, 34, 31, 27, 4, 12, 35, 39, 41, 8, 6, 13, 14, 17, 19, 20.

This data can be grouped and shown in tabular form as follows:

Class interval	Frequency	Cumulative frequency
0–6	10	10
7–13	11	21
14–20	13	34
21–27	4	38
28–34	7	45
35–41	10	55
42–48	5	60

Thus, the 60 values have been put into only 7 groups, called the classes. The width of the class is called the *class interval* and the number in that interval is called the *frequency*. The mid-point or the mid-value of the class is called the *class mark*.

8.4 FREQUENCY DISTRIBUTION

A table in which the frequencies and the associated values of a variable are written side by side, is known as a *frequency distribution*. A frequency distribution can be discrete or continuous depending upon whether the variable is discrete or continuous. A frequency distribution has the following parameters:

(i) Number of class intervals
(ii) Width of a class interval
(iii) Mid-value of a class
(iv) Cumulative frequency

8.4.1 Class Intervals

The class intervals can be exclusive or inclusive. In the exclusive class interval, the upper limit of a class is taken to be equal to the lower limit of the next class. To keep various class intervals as mutually exclusive, the observations with magnitude greater than or equal to lower limit but less than the upper limit of a class are included in it. For example, if the lower limit of a class is 20 and its upper limit is 30 then this class, written as 20–30, includes all the observations which are greater than or equal to 20 but less than 30. The observations with magnitude 30 will be included in the next class.

Class intervals	Frequency
0–10	5
10–20	17
20–30	25
30–40	12
40–50	8

In the inclusive class interval, all the observations with magnitude greater than or equal to lower limit and less than or equal to upper limit of a class are included in it.

Class intervals	Frequency
0–9	12
10–19	9
20–29	18
30–39	35
40–49	20

Inclusive class intervals can be converted into exclusive class intervals by the following procedures:

(i) Find the difference between the lower limit of the second class and the upper limit of the first class.

(ii) Divide the difference by 2.

(iii) Subtract the value so obtained from all the lower limits and add the value to all the uppper limits.

In the above example, the lower limit of the second class is 10 and the upper limit of the first class is 9. Hence, $\frac{10-9}{2}=0.5$ is subtracted from all the lower limits and added to all the upper limits as follows:

Class intervals	Frequency
−0.5–9.5	12
9.5–19.5	9
19.5–29.5	18
29.5–39.5	35
39.5–49.5	20

8.4.2 Mid-value of a Class

In exclusive types of class intervals, the mid-value of a class is defined as the arithmetic mean of its lower and upper limits.

8.4.3 Cumulative Frequency

There are two types of cumulative frequency distributions:

(i) *Less than cumulative frequency*: Less than cumulative frequency for any value of the variable/class is obtained by adding successively the frequencies of all the previous classes, including the frequency of the class, against which the total are written provided the values are written in ascending order of magnitude.

(ii) *More (or greater) than cumulative frequency*: More than cumulative frequency for any value of the variable/class is obtained by adding successively the frequencies of all the succeeding classes, including the frequency of the class, against which the total are written provided values are written in ascending order of magnitude.

8.5 GRAPHICAL REPRESENTATION

A frequency distribution is conveniently represented by means of a graph. Graphs are good visual aids. It makes the raw data readily intelligible and leaves a more lasting impression on the mind of the observer. But it does not give accurate measurements of the variable as are given by the table. Some important types of graphs are given below:

1. Histogram A histogram is drawn by erecting rectangles over the class intervals, such that the areas of the rectangles are proportional to the class frequencies. If the class intervals are of equal size, the height of the rectangles will be proportional to the class frequencies. For drawing a histogram, all the class intervals are marked off along the x-axis on a suitable scale and frequencies are marked off along the y-axis on a suitable scale. If, however, the classes are of unequal width then the height of the rectangle will be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called a histogram. If the grouped frequency distribution is not continuous, first it is to be converted into a continuous distribution and then the histogram is drawn. The frequency distribution and corresponding histogram are shown below:

Class intervals	Frequency
30–42	7
42–54	4
54–66	8
66–78	9
78–90	5
90–102	5
102–114	2

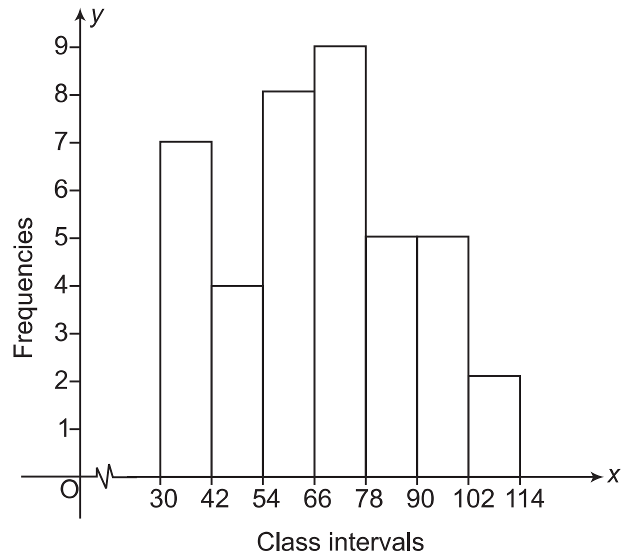

Fig. 8.1

2. Frequency Polygon A frequency polygon for an ungrouped frequency distribution is obtained by joining points plotted with the variable values as abscissae and the frequencies as the ordinates. For a grouped frequency distribution, the abscissae of the points are mid-values of the class intervals. For equal class intervals, the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by straight lines. If the class intervals are of small width, the polygon can be obtained by drawing a smooth curve through the vertices of the frequency polygon and is called the frequency curve.

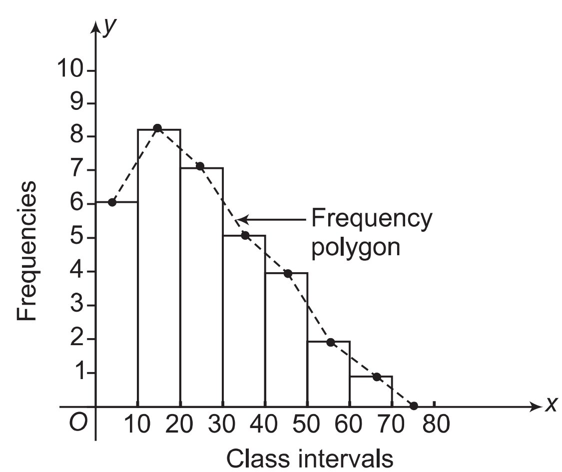

Fig. 8.2

3. Cumulative Frequency Curve or Ogive A cumulative frequency curve or ogive is obtained by plotting cumulative frequencies above or below a given value. Since a cumulative frequency distribution can be of 'less than' or 'more than' type and accordingly there are two types of ogives—'less than' ogive or 'more than' ogive.

A 'less than' ogive is obtained by plotting the points with the upper limits of the classes as abscissae and the corresponding less than cumulative frequency as ordinates and joining these points by a freehand smooth curve. A 'more than' ogive is obtained by plotting the points with the lower limits of the classes as abscissae and the corresponding more than cumulative frequency as ordinates and joining these points by a freehand smooth curve.

An ogive is used to determine certain positional averages like median, quartiles, deciles, percentiles, etc. Various frequency distributions can be compared on the basis of their ogives.

Example 1

Draw a histogram and frequency curve for the following data:

Profit (₹ in thousands)	0–15	15–30	30–45	45–60	60–75	75–90	90–105	105–120	120–135
No. of companies	3	7	18	25	20	12	6	5	2

Solution

Histogram and Frequency curve

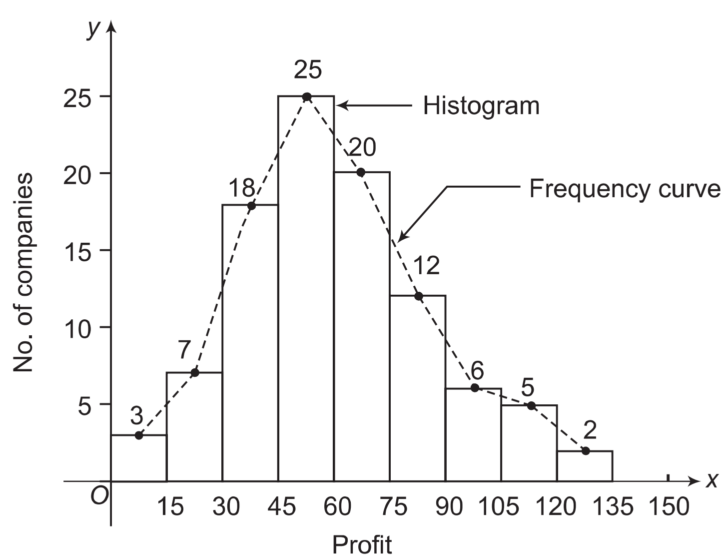

Fig. 8.3

Example 2

Draw a histogram and a frequency polygon for the following data:

Output (units per worker)	500–509	510–519	520–529	530–539	540–549	550–559	560–569
No. of workers	8	18	23	37	47	26	16

Solution

The data is presented in the form of inclusive class intervals. It can be converted into exclusive class intervals. The difference between the lower limit of the second class interval and the upper limit of the first class interval is 510 – 509 = 1. The new classes will be formed by subtracting $\frac{1}{2}$ from the lower limit and adding $\frac{1}{2}$ to the upper limit.

Class intervals	No. of workers (frequency)
499.5–509.5	8
509.5–519.5	18
519.5–529.5	23
529.5–539.5	37
539.5–549.5	47
549.5–559.5	26
559.5–569.5	16

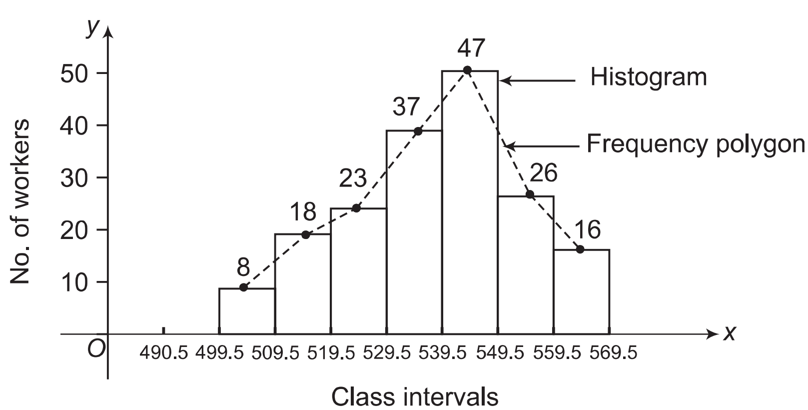

Fig. 8.4

Example 3

Construct a histogram and a frequency polygon for the following frequency distribution:

Marks (mid-value)	100	120	140	160	180	200
No. of students	5	6	4	6	4	5

Solution

The given data of mid-points is first converted into class interval form. The difference between two mid-values is 20. Hence, $\frac{20}{2}$ is subtracted from each mid-value to get the lower limit and $\frac{20}{2}$ is added to each mid-value to get the upper limit of a class interval.

Class intervals	No. of students
90–110	5
110–130	6
130–150	4
150–170	6
170–190	4
190–210	5

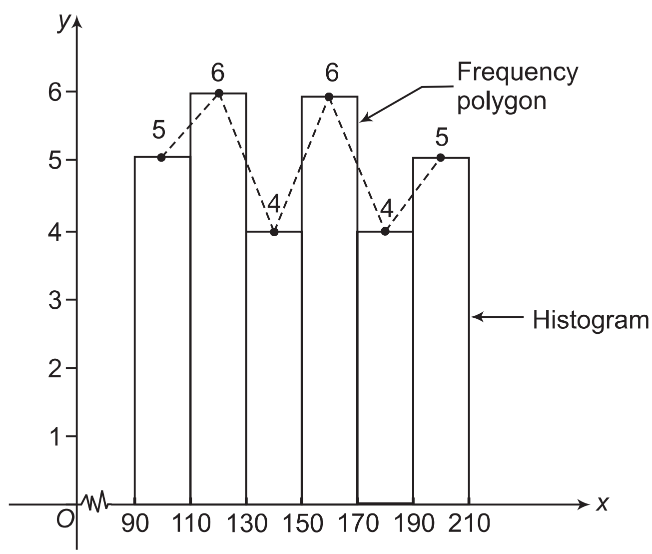

Fig. 8.5

Example 4

The following are the scores of two groups of a class in a test of reading ability:

Scores	Group A	Group B
50–52	4	2
47–49	10	3
44–46	15	4
41–43	18	8
38–40	20	12
35–37	12	17
32–34	13	22

Construct a frequency polygon for each group on the same axes.

Solution

For both the groups, i.e., group *A* and group *B*, the two hypothetical intervals with zero frequencies, one at the beginning and the other at the end with frequencies zero (53–55) and (29–31) are created.

Table for Group A

Scores	Class marks	Frequency	Points
53–55	54	0	(54, 0)
50–52	51	4	(51, 4)
47–49	48	10	(48, 10)
44–46	45	15	(45, 15)
41–43	42	18	(42, 18)
38–40	39	20	(39, 20)
35–37	36	12	(36, 12)
32–34	33	13	(33, 13)
29–31	30	0	(30, 0)

Table for Group B

Scores	Class marks	Frequency	Points
53–55	54	0	(54, 0)
50–52	51	2	(51, 2)
47–49	48	3	(48, 3)
44–46	45	4	(45, 4)
41–43	42	8	(42, 8)
38–40	39	12	(39, 12)
35–37	36	17	(36, 17)
32–34	33	22	(33, 22)
29–31	30	0	(30, 0)

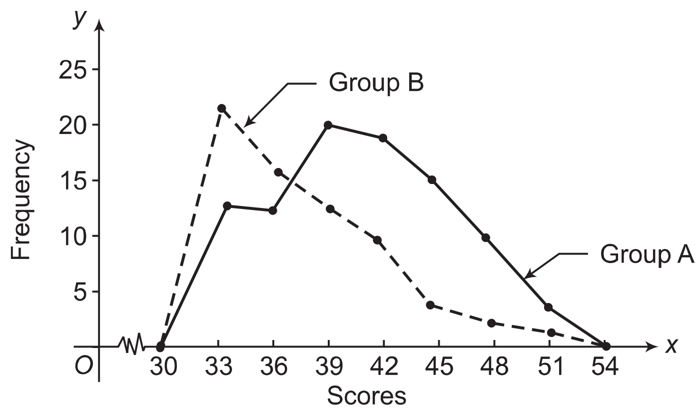

Fig. 8.6

Example 5

Draw 'less than' and 'more than' ogive distributions of monthly salary of 250 *families.*

Income intervals	0–500	500–1000	1000–1500	1500–2000	2000–2500	2500–3000	3000–3500	3500–4000
No. of families	50	80	40	25	25	15	10	5

Solution

'Less than' and 'More than' Frequency Distributions

Income intervals	No. of families	Less than cumulative frequency	More than cumulative frequency
0–500	50	50	250
500–1000	80	130	200
1000–1500	40	170	120
1500–2000	25	195	80
2000–2500	25	220	55
2500–3000	15	235	30
3000–3500	10	245	5
3500–4000	5	250	5

A 'less than' ogive is obtained by plotting the points (500, 50), (1000, 130), (1500, 170), (2000, 195), (2500, 220), (3000, 235), (3500, 245), (4000, 250) and joining them by a freehand curve.

A 'more than' ogive is obtained by plotting the points (0, 250), (500, 200), (1000, 120), (1500, 80), (2000, 55), (2500, 30), (3000, 15), (3500, 5) and joining them by a freehand curve.

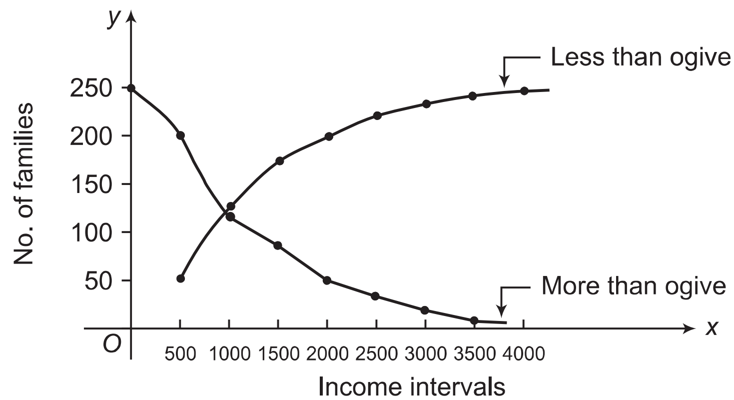

Fig. 8.7

Example 6

Draw the 'less than' ogive for the following distribution:

Age (in years)	0–9	10–19	20–29	30–39	40–49	50–59	60–69
No. of persons	5	15	20	25	15	12	8

Solution

The given frequency distribution is not continuous. It is first converted into continuous or exclusive class intervals.

Age (in years)	Class intervals	No. of persons	Cumulative frequency
0–9	–0.5–9.5	5	5
10–19	9.5–19.5	15	20
20–29	19.5–29.5	20	40
30–39	29.5–39.5	25	65
40–49	39.5–49.5	15	80
50–59	49.5–59.5	12	92
60–69	59.5–69.5	8	100

A 'less than' ogive is obtained by plotting points (9.5, 5), (19.5, 20), (29.5, 40), (39.5, 65), (49.5, 80), (59.5, 92), (69.5, 100) and joining them by a freehand smooth curve.

Fig. 8.8

Example 7

Convert the following distribution into a more than frequency distribution:

Weekly wages less than (₹)	20	40	60	80	100
No. of workers	41	92	156	194	201

For the data given, draw 'less than' and 'more than' ogives.

Solution

'Less than' and 'more than' frequency distribution.

Weekly wages (₹)	No. of workers *f*	Less than cumulative frequency	More than cumulative frequency
0–20	41	41	201
20–40	92–41 = 51	92	160
40–60	156–92 = 64	156	109
60–80	194–156 = 38	194	45
80–100	201–194 = 7	201	7

A 'less than' ogive is obtained by plotting the points (20, 41), (40, 92), (60, 156), (80, 194), (100, 201) and joining them by a freehand curve.

A 'more than' ogive is obtained by plotting the points (0, 201), (20, 160), (40, 109), (60, 45), (80, 7) and joining them by a freehand curve.

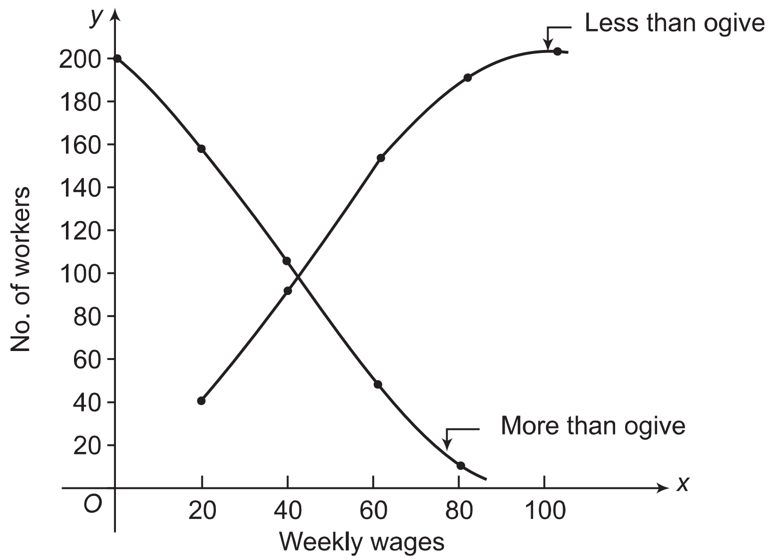

Fig. 8.9

EXERCISE 8.1

1. The following are the monthly rents in rupees of 40 shops. Tabulate the data by grouping in intervals of ₹ 8.

   38, 42, 49, 35, 82, 35, 77, 60, 50, 75, 84, 75, 63, 40, 70, 42, 36, 65, 51, 48, 74, 47, 50, 55, 64, 67, 72, 77, 82, 51, 31, 38, 43, 75, 67, 70, 43, 64, 84, 71.

2. The following table shows the distribution of the number of students per teacher in 750 colleges:

Students	1	4	7	10	13	16	19	22	25	28
Frequency	7	46	165	195	189	89	28	19	9	3

3. Draw a histogram for the following data:

Age (in years)	2–5	5–11	11–12	12–14	14–15	15–16
No. of boys	6	6	2	5	1	3

4. Draw the histogram and frequency polygon for the following data:

Monthly wages (₹ in thousands)	11–13	13–15	15–17	17–19	19–21	21–23	23–25
No. of workers	6	53	85	56	21	16	8

5. Draw the histogram and frequency polygon for the following distribution:

Class interval	0-99	100-199	200-299	300-399	400-499	500-599	600-699	700-799
Frequency	10	54	184	264	246	40	1	1

6. Represent the following distribution by (i) histogram and (ii) frequency polygon:

Scores	Frequency
30–39	1
40–49	3
50–59	14
60–69	20
70–79	22
80–89	12
90–99	2

7. Represent the following distribution by an ogive:

Marks	No. of students	Marks	No. of students
0–10	5	50–60	4
10–20	13	60–70	1
20–30	12	70–80	3
30–40	11	80–90	1
40–50	8	90–100	2

8. The following table gives the distribution of monthly income of 600 middle-class families in a certain city:

Monthly income in ₹	Frequency	Monthly income in ₹	Frequency
Below 76	69	300–375	58
76–150	167	375–450	25
150–225	207	450 and over	10
225–300	65		

Draw 'less than' and 'more than' ogive for the above data.

9. Draw an ogive by less than method for the following data:

No. of rooms	1	2	3	4	5	6	7	8	9	10
No. of houses	4	9	22	28	24	12	8	6	5	2

10. Draw histogram, frequency polygon and ogive for the following data:

Marks	Frequency	Marks	Frequency
0-10	4	40-50	20
10-20	10	50-60	18
20-30	16	60-70	8
30-40	22	70-80	2

8.6 MEASURES OF CENTRAL TENDENCY

Summarization of data is a necessary function of any statistical analysis. The data is summarized in the form of tables and frequency distributions. In order to bring the characteristics of the data, these tables and frequency distributions need to be summarized further. A measure of central tendency or an average is very essential and an important summary measure in any statistical analysis.

An *average* is a single value which can be taken as a representative of the whole distribution. There are five types of measures of central tendency or averages which are commonly used.

(i) Arithmetic mean
(ii) Median
(iii) Mode
(iv) Geometric mean
(v) Harmonic mean

A good measure of average must have the following characteristics:

(i) It should be rigidly defined so that different persons obtain the same value for a given set of data.
(ii) It should be easy to understand and easy to calculate.
(iii) It should be based on all the observations of the data.
(iv) It should be easily subjected to further mathematical calculations.
(v) It should not be much affected by the fluctuations of sampling.
(vi) It should not be unduly affected by extreme observations.
(vii) It should be easy to interpret.

8.7 ARITHMETIC MEAN

The *arithmetic mean* of a set of observations is their sum divided by the number of observations. Let $x_1, x_2, ..., x_n$ be n observations. Then their average or arithmetic mean is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}$$

For example, the marks obtained by 10 students in Class XII in a physics examination are 25, 30, 21, 55, 40, 45, 17, 48, 35, 42. The arithmetic mean of the marks is given by

$$\bar{x} = \frac{\sum x}{n} = \frac{25+30+21+55+40+45+17+48+35+42}{10} = \frac{358}{10} = 35.8$$

If n observations consist of n distinct values denoted by $x_1, x_2, ..., x_n$ of the observed variable x occurring with frequencies $f_1, f_2, ..., f_n$ respectively then the arithmetic mean is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_nx_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_ix_i}{N} = \frac{\sum fx}{N}$$

where $N = \sum_{i=1}^{n} f_i = f_1 + f_2 + \cdots + f_n$

8.7.1 Arithmetic Mean of Grouped Data

In case of a grouped or continuous frequency distribution, the arithmetic mean is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i} = \frac{\sum fx}{N}, \text{ where } N = \sum_{i=1}^{n} f_i$$

and x is taken as the mid-value of the corresponding class.

Example 1

Find the arithmetic mean from the following frequency distribution:

x	5	6	7	8	9	10	11	12	13	14
f	25	45	90	165	112	96	81	26	18	12

Solution

x	f	fx
5	25	125
6	45	270
7	90	630
8	165	1320
9	112	1008
10	96	960
11	81	891
12	26	312
13	18	234
14	12	168
	$\sum f = 670$	$\sum fx = 5918$

$$N = \sum f = 670$$

$$\bar{x} = \frac{\sum fx}{N} = \frac{5918}{670} = 8.83$$

Example 2

Find the arithmetic mean of the marks from the following data:

Marks	0–10	10–20	20–30	30–40	40–50	50–60
Number of students	12	18	27	20	15	8

Solution

Marks	Number of students *(f)*	Mid-value *(x)*	*fx*
0–10	12	5	60
10–20	18	15	270
20–30	27	25	675
30–40	20	35	700
40–50	15	45	675
50–60	8	55	440
	$\sum f = 100$		$\sum fx = 2820$

$$N = \sum f = 100$$

$$\bar{x} = \frac{\sum fx}{N} = \frac{2820}{100} = 28.20$$

8.7.2 Arithmetic Mean from Assumed Mean

If the values of x and (or) f are large, the calculation of mean becomes quite time-consuming and tedious. In such cases, the provisional mean 'a' is taken as that value of x (mid-value of the class interval) which corresponds to the highest frequency or which comes near the middle value of the frequency distribution. This number is called the *assumed mean*.

Let $$d = x - a$$

$$fd = f(x-a) = fx - af$$

$$\sum fd = \sum fx - a\sum f$$

$$= \sum fx - aN$$

Dividing both the sides by n,

$$\frac{\sum fd}{N} = \frac{\sum fx}{N} - a$$
$$= \bar{x} - a$$
$$\therefore \quad \bar{x} = a + \frac{\sum fd}{N}$$

Example 1

Ten coins were tossed together and the number of tails resulting from them were observed. The operation was performed 1050 times and the frequencies thus obtained for different numbers of tail (x) are shown in the following table. Calculate the arithmetic mean.

x	0	1	2	3	4	5	6	7	8	9	10
f	2	8	43	133	207	260	213	120	54	9	1

Solution

Let $a = 5$ be the assumed mean.

$d = x - a = x - 5$

x	f	$d = x - 5$	fd
0	2	–5	–10
1	8	–4	–32
2	43	–3	–129
3	133	–2	–266
4	207	–1	–207
5	260	0	0
6	213	1	213
7	120	2	240
8	54	3	162
9	9	4	36
10	1	5	5
	$\sum f = 1050$		$\sum fd = 12$

$$N = \sum f = 1050$$
$$\bar{x} = a + \frac{\sum fd}{N}$$

$$= 5 + \frac{12}{1050}$$
$$= 5.0114$$

Example 2

Calculate the mean for the following frequency distribution

Class	0–8	8–16	16–24	24–32	32–40	40–48
Frequency	8	7	16	24	15	7

Solution

Let $a = 28$ be the assumed mean.
$d = x - a = x - 28$

Class	Frequency	Mid-value (x)	$d = x - 28$	fd
0–8	8	4	–24	–192
8–16	7	12	–16	–112
16–24	16	20	–8	–128
24–32	24	28	0	0
32–40	15	36	8	120
40–48	7	44	16	112
	$\sum f = 77$			$\sum fd = -200$

$$N = \sum f = 77$$
$$\overline{x} = a + \frac{\sum fd}{N}$$
$$= 28 + \frac{(-200)}{77}$$
$$= 25 \cdot 403$$

8.7.3 Arithmetic Mean by the Step-Deviation Method

When the class intervals in a grouped data are equal, calculation can be simplified by the step-deviation method. In such cases, deviation of the variate x from the assumed mean a (i.e., $d = x - a$) are divided by the common factor h which is equal to the width of the class interval.

Let $d = \dfrac{x - a}{h}$

$$\overline{x} = a + h\frac{\sum fd}{\sum f} = a + h\frac{\sum fd}{N}$$

where a is the assumed mean

$d = \frac{x-a}{h}$ is the deviation of any variate x from a

h is the width of the class interval

N is the number of observations

Example 1

Calculate the arithmetic mean of the following marks obtained by students in mathematics:

Marks (x)	5	10	15	20	25	30	35	40	45	50
Number of students (f)	20	43	75	67	72	45	39	9	8	6

Solution

Let $a = 30$ be the assumed mean and $h = 5$ be the width of the class interval.

$$d = \frac{x-a}{h} = \frac{x-30}{5}$$

x	f	$d = \frac{x-30}{5}$	fd
5	20	–5	–100
10	43	–4	–172
15	75	–3	–225
20	67	–2	–134
25	72	–1	–72
30	45	0	0
35	39	1	39
40	9	2	18
45	8	3	24
50	6	4	24
	$\sum f = 384$		$\sum fd = -598$

$$N = \sum f = 384$$

$$\begin{aligned} \bar{x} &= a + h\frac{\sum fd}{N} \\ &= 30 + 5\left(\frac{-598}{384}\right) \\ &= 22.214 \end{aligned}$$

Example 2

The following table gives the distribution of companies according to size of capital. Find the mean size of the capital of a company.

Capital (₹ in lacs)	<5	<10	<15	<20	<25	<30
No. of companies	20	27	29	38	48	53

Solution

This is a 'less than' type of frequency distribution. This will be first converted into class intervals. Let $a = 12.5$ be the assumed mean and $h = 5$ be the width of the class interval.

$$d = \frac{x-a}{h} = \frac{x-12.5}{5}$$

Class intervals	Frequency f	Mid-value x	$d = \dfrac{x-12.5}{5}$	fd
0–5	20	2.5	–2	–40
5–10	7	7.5	–1	–7
10–15	2	12.5	0	0
15–20	9	17.5	1	9
20–25	10	22.5	2	20
25–30	5	27.5	3	15
	$\sum f = 53$			$\sum fd = -3$

$$N = \sum f = 53$$

$$\bar{x} = a + h\frac{\sum fd}{N}$$

$$= 12.5 + 5\left(\frac{-3}{53}\right)$$

$$= 12.22 \text{ lacs}$$

Example 3

Following is the distribution of marks obtained by 60 *students in a mathematics test:*

Marks	Number of students
More than 0	60
More than 10	56
More than 20	40
More than 30	20
More than 40	10
More than 50	3

Calculate the arithmetic mean.

Solution

This is a 'more than' type of frequency distribution. This will be first converted into class intervals. Let a = 35 be the assumed mean and h = 10 be the width of the class interval.

$$d = \frac{x - a}{h} = \frac{x - 35}{10}$$

Marks	No. of students f	Mid-value x	$d = \frac{x-35}{10}$	fd
0–10	4	5	–3	–12
10–20	16	15	–2	–32
20–30	20	25	–1	–20
30–40	10	35	0	0
40–50	7	45	1	7
50–60	3	55	2	6
	$\sum f = 60$			$\sum fd = -51$

$$N = \sum f = 60$$

$$\begin{aligned}\bar{x} &= a + h\frac{\sum fd}{N} \\ &= 35 + 10\left(\frac{-51}{60}\right) \\ &= 26.5\end{aligned}$$

EXERCISE 8.2

1. Find the mean of the following marks obtained by students of a class:

Marks	15	20	25	30	35	40
No. of students	9	7	12	14	15	6

[**Ans.:** 25.58]

2. The following table gives the distribution of total household expenditure (in rupees) of manual workers in a city:

Expenditure (in ₹)	100–150	150–200	200–250	250–300	300–350	350–400	400–450	450–500
Frequency	24	40	33	28	30	22	16	7

Find the average expenditure (in ₹) per household.

[**Ans.:** ₹ 266.25]

3. Calculate the mean for the following data:

Heights (in cm)	135–140	140–145	145–150	150–155	155–160	160–165	165–170	170–175
No. of boys	4	9	18	28	24	10	5	2

[**Ans.:** 153.45 cm]

4. The weights in kilograms of 60 workers in a factory are given below. Find the mean weight of a worker.

Weight (in kg)	60	61	62	63	64	65
No. of workers	5	8	14	16	10	7

[**Ans.:** 62.65 kg]

5. Calculate the mean from the following data:

Marks less than/up to	10	20	30	40	50	60
No. of students	10	30	60	110	150	180

[**Ans.:** 35]

6. Calculate the mean from the following data:

Marks more than	0	10	20	30	40	50	60
No. of students	180	170	150	120	70	30	0

[**Ans.:** 35]

7. Calculate the mean from the following data:

Marks	1–5	6–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45
No. of students	7	10	16	30	24	17	10	5	1

[**Ans.:** 20.33]

8.8 MEDIAN

Median is the central value of the variable when the values are arranged in ascending or descending order of magnitude. It divides the distribution into two equal parts. When the observations are arranged in the order of their size, median is the value of that item which has equal number of observations on either side.

In case of ungrouped data, if the number of observations is odd then the median is the middle value after the values have been arranged in ascending or descending order of magnitude. If the number of observations is even, there are two middle terms and the median is obtained by taking the arithmetic mean of the middle terms.

Examples

(i) The median of the values 20, 15, 25, 28, 18, 16, 30, i.e., 15, 16, 18, 20, 25, 28, 30 is 20 because $n = 7$, i.e., odd and the median is the middle value, i.e., 20.

(ii) The median of the values 8, 20, 50, 25, 15, 30, i.e., 8, 15, 20, 25, 30, 50 is the arithmetic mean of the middle terms, i.e., $\frac{20+25}{2} = 22.5$ because $n = 6$, i.e., even.

In case of discrete frequency distribution, the median is obtained by considering the cumulative frequencies. The steps for calculating the median are given below:

(i) Arrange the values of the variables in ascending or descending order of magnitudes.

(ii) Find $\frac{N}{2}$, where $N = \sum f$

(iii) Find the cumulative frequency just greater than $\frac{N}{2}$ and determine the corresponding value of the variable.

(iv) The corresponding value of x is the median.

Example 1

The following table represents the marks obtained by a batch of 12 *students in certain class tests in physics and chemistry.*

Marks (Physics)	53	54	32	30	60	46	28	25	48	72	33	65
Marks (Chemistry)	55	41	48	49	27	25	23	20	28	60	43	67

Indicate the subject in which the level of achievement is higher.

Solution

The level of achievement is higher in that subject for which the median marks are more.

Arranging the marks in two subjects in ascending order,

Marks (Physics)	25	28	30	32	33	46	48	53	54	60	65	72
Marks (Chemistry)	20	23	25	27	28	41	43	48	49	55	60	67

Since the number of students is 12, the median is the arithmetic mean of the middle terms.

Median marks in physics = $\dfrac{46+48}{2} = 47$

Median marks in chemistry = $\dfrac{41+43}{2} = 42$

Since the median marks in physics are greater than the median marks in chemistry, the level of achievement is higher in physics.

Example 2

Obtain the median for the following frequency distribution.

x	0	1	2	3	4	5	6	7
f	7	14	18	36	51	54	52	18

Solution

x	f	Cumulative frequency
0	7	7
1	14	21
2	18	39
3	36	75
4	51	126
5	54	180
6	52	232
7	18	250

$N = 250$

$$\frac{N}{2} = \frac{250}{2} = 125$$

The cumulative frequency just greater than $\dfrac{N}{2}$ = 125 is 126 and the value of x corresponding to 126 is 4. Hence, the median is 4.

Median for Continuous Frequency Distribution

In case of continuous frequency distribution (less than frequency distribution), the class corresponding to the cumulative frequency just greater than $\frac{N}{2}$, is called the *median class*, and the value of the median is given by

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

where l is the lower limit of the median class
f is the frequency of the median class
h is the width of the median class
c is the cumulative frequency of the class preceding the median class
N is sum of frequencies, i.e., $N = \sum f$

In case of 'more than' or 'greater than' type of frequency distributions, the value of the median is given by

$$\text{Median} = u - \frac{h}{f}\left(\frac{N}{2} - c\right)$$

where u is the upper limit of the median class
f is the frequency of the median class
h is the width of the median class
c is the cumulative frequency of the class succeeding the median class

Example 1

The following table gives the weekly expenditures of 100 *workers. Find the median weekly expenditure.*

Weekly expenditure (in ₹)	0–10	10–20	20–30	30–40	40–50
Number of workers	14	23	27	21	15

Solution

Weekly expenditure (in ₹)	Number of workers (f)	Cumulative frequency
0–10	14	14
10–20	23	37
20–30	27	64
30–40	21	85
40–50	15	100

$N = 100$

$$\frac{N}{2} = \frac{100}{2} = 50$$

The cumulative frequency just greater than $\frac{N}{2} = 50$ is 64 and the corresponding class 20–30 is the median class.

Here, $\frac{N}{2} = 50, \quad l = 20, \quad h = 10, \quad f = 27, \quad c = 37$

$$\begin{aligned} \text{Median } &= l + \frac{h}{f}\left(\frac{N}{2} - c\right) \\ &= 20 + \frac{10}{27}(50 - 37) \\ &= 24.815 \end{aligned}$$

Example 2

From the following data, calculate the median:

Marks (Less than)	5	10	15	20	25	30	35	40	45
No. of students	29	224	465	582	634	644	650	653	655

[Summer 2015]

Solution

This is a 'less than' type of frequency distribution. This will be first converted into class intervals.

Class intervals	Frequency	Less than CF
0–5	29	29
5–10	195	224
10–15	241	465
15–20	117	582
20–25	52	634
25–30	10	644
30–35	6	650
35–40	3	653
40–45	2	655

$$N = 655$$

Since $\frac{N}{2} = \frac{655}{2} = 327.5$, the median class is 10–15.

Here, $l = 10, \quad h = 5, \quad f = 241, \quad c = 224$

$$\begin{aligned}\text{Median} &= l + \frac{h}{f}\left(\frac{N}{2} - c\right) \\ &= 10 + \frac{5}{241}(327.5 - 224) \\ &= 12.147\end{aligned}$$

Example 3

Find the mean of the following data:

Age greater than (in years)	0	10	20	30	40	50	60	70
No. of persons	230	218	200	165	123	73	28	8

Solution

This is a 'greater than' type of frequency distribution. This will be first converted into class intervals.

Class intervals	Frequency	Greater than CF
0–10	12	230
10–20	18	218
20–30	35	200
30–40	42	165
40–50	50	123
50–60	45	73
60–70	20	28
70 and above	8	8

$N = 230$

Since $\frac{N}{2} = \frac{230}{2} = 115$, the median class is 40–50.

Here, $u = 50, \quad h = 10, \quad f = 50, \quad c = 73$

$$\begin{aligned}\text{Median} &= u - \frac{h}{f}\left(\frac{N}{2} - c\right) \\ &= 50 - \frac{10}{50}(115 - 73) \\ &= 41.6 \text{ years}\end{aligned}$$

Example 4

The following table gives the marks obtained by 50 students in mathematics. Find the median.

Marks	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49
No. of students	4	6	10	5	7	3	9	6

Solution

Since the class intervals are inclusive, it is necessary to convert them into exclusive series.

Marks	No. of students	Cumulative frequency
9.5–14.5	4	4
14.5–19.5	6	10
19.5–24.5	10	20
24.5–29.5	5	25
29.5–34.5	7	32
34.5–39.5	3	35
39.5–44.5	9	44
44.5–49.5	6	50

$N = 50$

Since $\frac{N}{2} = \frac{50}{2} = 25$, the median class is 24.5–29.5.

Here, $l = 24.5$, $h = 5$, $f = 5$, $c = 20$

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$= 24.5 + \frac{5}{5}(25 - 20)$$

$$= 29.5$$

Example 5

Find the median of the following distribution:

Mid-values	1500	2500	3500	4500	5500	6500	7500
Frequency	27	32	65	78	58	32	8

Solution

The difference between two mid-values is 1000. On subtracting and adding half of this, i.e., 500 to each of the mid-values, the lower and upper limits of the respective class intervals are obtained.

Class intervals	Frequency	Cumulative frequency
1000–2000	27	27
2000–3000	32	59
3000–4000	65	124
4000–5000	78	202
5000–6000	58	260
6000–7000	32	292
7000–8000	8	300

$N = 300$

Since $\frac{N}{2} = 150$, the median class is 4000–5000.

Here, $l = 4000$, $h = 1000$, $f = 78$, $c = 124$

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$= 4000 + \frac{1000}{78}(150 - 124)$$

$$= 4333.33$$

EXERCISE 8.3

1. The heights (in cm) of 15 students of Class XII are 152, 147, 156, 149, 151, 159, 148, 160, 153, 154, 150, 143, 155, 157, 161. Find the median.

[**Ans.:** 153 cm]

2. The median of the following observations are arranged in the ascending order: 11, 12, 14, 18, $x + 2$, $x + 4$, 30, 32, 35, 41 is 24. Find x.

[**Ans.:** 21]

3. Find the median of the following frequency distribution:

x	10	11	12	13	14	15	16
f	8	15	25	20	12	10	5

[**Ans.:** 12]

4. Find the median of the following frequency distribution:

Wages (in ₹)	20–30	30–40	40–50	50–60	60–70
No. of workers	3	5	20	10	5

[**Ans.:** 46.75]

5. Calculate the median of the following data:

x	3–4	4–5	5–6	6–7	7–8	8–9	9–10	10–11
f	3	7	12	16	22	20	13	7

[**Ans.:** 7.55]

6. The weekly wages of 1000 workers of a factory are shown in the following table:

Weekly wages (less than)	425	475	525	575	625	675	725	775	825	875
No. of workers	2	10	43	123	293	506	719	864	955	1000

[**Ans.:** 673.59]

7. Calculate the mean of the following distribution of marks obtained by 50 students in advanced engineering mathematics.

Marks more than	0	10	20	30	40	50
No. of students	50	46	40	20	10	3

[**Ans.:** 27.5]

8. Calculate the median from the following data:

Mid-values	115	125	135	145	155	165	175	185	195
Frequency	6	25	48	72	116	60	38	22	3

[**Ans.:** 153.79]

8.9 MODE

Mode is the value which occurs most frequently in a set of observations and around which the other items of the set are heavily distributed. In other words, mode is the value of the variable which is most frequent or predominant in the series. In case of a discrete frequency distribution, mode is the value of x corresponding to the maximum frequency.

Examples

(i) In the series 6, 5, 3, 4, 3, 7, 8, 5, 9, 5, 4, the value 5 occurs most frequently. Hence, the mode is 5.

(ii) Consider the following frequency distribution:

x	1	2	3	4	5	6	7	8
f	4	9	16	25	22	15	7	3

The value of x corresponding to the maximum frequency, viz., 25, is 4. Hence, the mode is 4.

For an asymmetrical frequency distribution, the difference between the mean and the mode is approximately three times the difference between the mean and the median.

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

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

This is known as the *empirical formula for calculation of the mode*.

Mode for a Continuous Frequency Distribution

In case of a continuous frequency distribution, the class in which the mode lies is called the *modal class* and the value of the mode is given by

$$\text{Mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

where l is the lower limit of the modal class

h is the width of the modal class

f_m is the frequency of the modal class

f_1 is the frequency of the class preceding the modal class

f_2 is the frequency of the class succeeding the modal class

This method of finding the mode is called the *method of interpolation*. This formula is applicable only to a unimodal frequency distribution.

Example 1

Find the mode for the following data:

Profit per shop	0–100	100–200	200–300	300–400	400–500	500–600
No. of shops	12	18	27	20	17	6

Solution

Since the maximum frequency is 27, which lies in the class 200–300, the modal class is 200–300.

Here, $l = 200$, $h = 100$, $f_m = 27$, $f_1 = 18$, $f_2 = 20$

$$\text{Mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

$$= 200 + 100\left[\frac{27 - 18}{2(27) - 18 - 20}\right]$$

$$= 256.25$$

Example 2

The frequency distribution of marks obtained by 60 *students of a class in a college is given by*

Marks	30–34	35–39	40–44	45–49	50–54	55–59	60–64
Frequency	3	5	12	18	14	6	2

Find the mode of the distribution.

Solution

The class intervals are first converted into a continuous exclusive series as shown in the following table:

Marks	Frequency
29.5–34.5	3
34.5–39.5	5
39.5–44.5	12
44.5–49.5	18
49.5–54.5	14
54.5–59.5	6
59.5–64.5	2

Since the maximum frequency is 18 which lies in the interval 44.5–49.5, the modal class is 44.5–49.5.

Here, $l = 44.5$, $h = 5$, $f_m = 18$, $f_1 = 12$, $f_2 = 14$

$$\text{Mode} = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

$$= 44.5 + 5\left[\frac{18 - 12}{2(18) - 12 - 14}\right]$$

$$= 47.5$$

Example 3

Find the mode for the following distribution:

Class intervals	0–10	10–20	20–30	30–40	40–50
Frequency	45	20	14	7	3

Solution

Since the highest frequency occurs in the first class interval, the interpolation formula is not applicable. Thus, empirical formula is used for calculation of mode.

Class intervals	Frequency	CF	Mid-value	$d = \frac{x-25}{10}$	fd
0–10	45	45	5	–2	–90
10–20	20	65	15	–1	–20
20–30	14	79	25	0	0
30–40	7	86	35	1	7
40–50	3	89	45	2	6
	$\sum f = 89$				$\sum fd = -97$

$$N = \sum f = 89$$

Since $\frac{N}{2} = \frac{89}{2} = 44.5$, the median class is 0–10.

Here, $l = 0$, $h = 10$, $f = 45$, $c = 0$

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$= 0 + \frac{10}{45}(44.5 - 0)$$

$$= 9.89$$

$$\text{Mean} = a + h\frac{\sum fd}{N}$$

$$= 25 + 10\left(\frac{-97}{89}\right)$$

$$= 14.1$$

$$\text{Hence, mode} = 3\,\text{Median} - 2\,\text{Mean}$$

$$= 3(9.89) - 2(14.1)$$

$$= 1.47$$

EXERCISE 8.4

1. Calculate the mode for the following distribution:

x	6	12	18	24	30	36
f	12	24	36	38	37	6

[**Ans.:** 24]

2. Calculate the mode for the following distribution:

x	10	20	30	40	50	60	70
f	17	22	31	39	27	15	13

[**Ans.:** 40]

3. Calculate the mode for the following distribution:

Class interval	0–4	4–8	8–12	12–16
Frequency	4	8	5	6

[**Ans.:** 6.28]

4. Calculate the mode of the following distribution:

x	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45
f	20	24	32	28	20	16	37	10	18

[**Ans.:** 13.33]

5. Calculate the mode for the following data:

Class	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	24	42	56	66	108	130	154

[**Ans.:** 71.348]

6. Find the mode of the following distribution:

Class	55–64	65–74	75–84	85–94	95–104	105–114	115–124	125–134	135–144
f	1	2	9	22	33	22	8	2	1

[**Ans.:** 99.5]

7. Calculate the modal marks from the following distribution of marks of 100 students of a class:

Marks (more than)	90	80	70	60	50	40	30	20	10
No. of students	0	4	15	33	53	76	92	98	100

[Ans.: 47]

8.10 STANDARD DEVIATION

Standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by the Greek letter σ. Let X be a random variable which takes on values, viz., $x_1, x_2, ..., x_n$. The standard deviation of these n observations is given by

$$\sigma = \sqrt{\frac{\sum (x-\overline{x})^2}{n}}$$

where $\overline{x} = \dfrac{\sum x}{n}$ is the arithmetic mean of these observations.

This equation can be modified further.

$$\begin{aligned}
\sigma &= \sqrt{\frac{\sum (x^2 - 2x\,\overline{x} + \overline{x})^2}{n}} \\
&= \sqrt{\frac{\sum x^2 - 2\overline{x}\sum x + \overline{x}^2 \sum 1}{n}} \\
&= \sqrt{\frac{\sum x^2}{n} - 2\frac{\sum x}{n}\frac{\sum x}{n} + \left(\frac{\sum x}{n}\right)^2 \cdot \frac{n}{n}} \qquad \left[\because \ \sum 1 = n\right] \\
&= \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \\
&= \sqrt{\text{Mean of squares} - \text{Square of mean}}
\end{aligned}$$

In case of a frequency distribution consisting of n observations $x_1, x_2, ..., x_n$ with respective frequencies $f_1, f_2, ..., f_n$, the standard deviation is given by

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

This equation can also be modified.

$$\sigma = \sqrt{\frac{\sum f\,(x^2 - 2x\bar{x} + \bar{x}^2)}{N}}$$

$$= \sqrt{\frac{\sum fx^2}{N} - \frac{2\bar{x}\sum fx}{N} + \bar{x}^2\frac{\sum f}{N}}$$

$$= \sqrt{\frac{\sum fx^2}{N} - 2\frac{\sum fx}{N}\frac{\sum fx}{N} + \left(\frac{\sum fx}{N}\right)^2} \qquad \left[\because\ \sum f = N \quad \text{and} \quad \bar{x} = \frac{\sum fx}{N}\right]$$

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

8.10.1 Variance

The *variance* is the square of the standard deviation and is denoted by σ^2. The method for calculating variance is same as that given for the standard deviation.

Example 1

Calculate the standard deviation of the weights of ten persons.

Weight (in kg)	45	49	55	50	41	44	60	58	53	55

Solution

$$n = 10$$

$$\sum x = 45 + 49 + 55 + 50 + 41 + 44 + 60 + 58 + 53 + 55 = 510$$

$$\sum x^2 = 45^2 + 49^2 + 55^2 + 50^2 + 41^2 + 44^2 + 60^2 + 58^2 + 53^2 + 55^2 = 26366$$

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

$$= \sqrt{\frac{26366}{10} - \left(\frac{510}{10}\right)^2}$$

$$= 5.967$$

Aliter:

$$\bar{x} = \frac{\sum x}{n} = \frac{510}{10} = 51$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
45	–6	36
49	–2	4
55	4	16
50	–1	1
41	–10	100
44	–7	49
60	9	81
58	7	49
53	2	4
55	4	16
		$\sum(x-\bar{x})^2 = 356$

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

$$= \sqrt{\frac{356}{10}}$$

$$= 5.967$$

Example 2

Calculate the standard deviation of the following data:

x	10	11	12	13	14	15	16	17	18
f	2	7	10	12	15	11	10	6	3

Solution

x	f	fx	x^2	fx^2
10	2	20	100	200
11	7	77	121	847
12	10	120	144	1440
13	12	156	169	2028
14	15	210	196	2940
15	11	165	225	2475
16	10	160	256	2560
17	6	102	289	1734
18	3	54	324	972
	$\sum f = 76$	$\sum fx = 1064$		$\sum fx^2 = 15196$

$$N = \sum f = 76$$

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

$$= \sqrt{\frac{15196}{76} - \left(\frac{1064}{76}\right)^2}$$

$$= 1.987$$

Aliter:

$$N = \sum f = 76$$

$$\bar{x} = \frac{\sum fx}{N} = \frac{1064}{76} = 14$$

x	f	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
10	2	–4	16	32
11	7	–3	9	63
12	10	–2	4	40
13	12	–1	1	12
14	15	0	0	0
15	11	1	1	11
16	10	2	4	40
17	6	3	9	54
18	3	4	16	48
				$\sum f(x-\bar{x})^2 = 300$

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

$$= \sqrt{\frac{300}{76}}$$

$$= 1.987$$

8.10.2 Standard Deviation from the Assumed Mean

If the values of x and f are large, the calculation of fx, fx^2 becomes tedious. In such a case, the assumed mean a is taken to simplify the calculation.

Let a be the assumed mean.

$$d = x - a$$

$$x = a + d$$

$$\sum fx = \sum f(a+d) = Na + \sum fd$$

Dividing both the sides by N,

$$\frac{\sum fx}{N} = a + \frac{\sum fd}{N}$$

$$\bar{x} = a + \bar{d}$$

$$x - \bar{x} = d - \bar{d}$$

$$\sigma_x = \sqrt{\frac{\sum f(x-\bar{x})^2}{N}}$$

$$= \sqrt{\frac{\sum f(d-\bar{d})^2}{N}}$$

$$= \sigma_d$$

Hence, the standard deviation is independent of change of origin.

$$\therefore \qquad \sigma_x = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Example 1

Find the standard deviation from the following data:

Size of the item	10	11	12	13	14	15	16
Frequency	2	7	11	15	10	4	1

Solution

Let $a = 13$ be the assumed mean.

$$d = x - a = x - 13$$

Size of item (x)	Frequency (f)	$d = x - a$	d^2	fd	fd^2
10	2	–3	9	–6	18
11	7	–2	4	–14	28
12	11	–1	1	–11	11
13	15	0	0	0	0
14	10	1	1	10	10
15	4	2	4	8	16
16	1	3	9	3	9
	$\sum f = 50$			$\sum fd = -10$	$\sum fd^2 = 92$

$$N = \sum f = 50$$

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

$$= \sqrt{\frac{92}{50} - \left(\frac{-10}{50}\right)^2}$$

$$= 1.342$$

8.10.3 Standard Deviation by Step-Deviation Method

Let a be the assumed mean and h be the width of the class interval.

$$d = \frac{x - a}{N}$$

$$x = a + hd$$

$$\sum fx = \sum f\ (a + hd) = Na + h\sum fd$$

Dividing both the sides by N,

$$\frac{\sum fx}{N} = a + h\frac{\sum fd}{N}$$

$$\bar{x} = a + h\bar{d}$$

$$x - \bar{x} = h(d - \bar{d})$$

$$\sigma_x = \sqrt{\frac{\sum f\,(x - \bar{x})^2}{N}}$$

$$= \sqrt{\frac{\sum f\ h^2\ (d - \bar{d})^2}{N}}$$

$$= h\sqrt{\frac{\sum f\ (d - \bar{d})^2}{N}}$$

$$= h\,\sigma_d$$

Hence, the standard deviation is independent of change of origin but not of scale.

$$\therefore \qquad \sigma_x = h\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Example 1

Find the standard deviation for the following distribution:

Marks	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Number of students	5	12	15	20	10	4	2

Solution

Let $a = 45$ be the assumed mean and $h = 10$ be the width of the class interval.

$$d = \frac{x-a}{h} = \frac{x-45}{10}$$

Marks	Number of students f	Mid-value x	$d = \frac{x-45}{10}$	d^2	fd	fd^2
10–20	5	15	–3	9	–15	45
20–30	12	25	–2	4	–24	48
30–40	15	35	–1	1	–15	15
40–50	20	45	0	0	0	0
50–60	10	55	1	1	10	10
60–70	4	65	2	4	8	16
70–80	2	75	3	9	6	18
	$\sum f = 68$				$\sum fd = -30$	$\sum fd^2 = 152$

$$N = \sum f = 68$$

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

$$= 10\sqrt{\frac{152}{68} - \left(\frac{-30}{68}\right)^2}$$

$$= 14.285$$

Example 2

Find the mean and standard deviation of the following distribution:

Age (in years)	No. of persons
less than 20	0
less than 25	170
less than 30	280
less than 35	360
less than 40	405
less than 45	445
less than 50	480

Solution

This is a 'less than' type of frequency distribution. This is first converted into an exclusive series. Let $a = 32.5$ be the assumed mean and $h = 5$ be the width of the class interval.

$$d = \frac{x-a}{h} = \frac{x-32.5}{5}$$

Class intervals	No. of persons f	Mid-value x	$d = \frac{x-32.5}{5}$	fd	fd^2
20–25	170	22.5	–2	–340	680
25–30	110	27.5	–1	–110	110
30–35	80	32.5	0	0	0
35–40	45	37.5	1	45	45
40–45	40	42.5	2	80	160
45–50	35	47.5	3	105	315
	$\sum f = 480$			$\sum fd = -220$	$\sum fd^2 = 1310$

$$N = \sum f = 480$$

$$\bar{x} = a + h\frac{\sum fd}{N}$$

$$= 32.5 + 5\left(\frac{-220}{480}\right)$$

$$= 30.21 \text{ years}$$

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

$$= 5\sqrt{\frac{1310}{480} - \left(\frac{-220}{480}\right)^2}$$

$$= 7.94 \text{ years}$$

8.10.4 Coefficient of Variation

The *standard deviation* is an absolute measure of dispersion. The coefficient of variation is a relative measure of dispersion and is denoted by CV.

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

where σ is the standard deviation and $\bar{x}$ is the mean of the given series. The coefficient of variation has great practical significance and is the best measure of comparing the variability of two series. The series or groups for which the coefficient of variation is greater is said to be more variable or less consistent. On the other hand, the series for which the variation is lesser is said to be less variable or more consistent.

Example 1

The arithmetic mean of the runs scored by three batsmen Amit, Sumeet, and Nayan in the series are 50, 48, *and* 12 *respectively. The standard deviations of their runs are* 15, 12, *and* 2 *respectively. Who is the more consistent of the three?*

Solution

Let $\bar{x}_1, \bar{x}_2, \bar{x}_3$ be the arithmetic means and $\sigma_1, \sigma_2, \sigma_3$ be the standard deviations of the runs scored by Amit, Sumeet, and Nayan.

$$\bar{x}_1 = 50, \bar{x}_2 = 48, \bar{x}_3 = 12, \sigma_1 = 15, \sigma_2 = 12, \sigma_3 = 2$$

$$\text{CV}_1 = \frac{\sigma_1}{\bar{x}_1} \times 100$$

$$= \frac{15}{50} \times 100$$

$$= 30\%$$

$$\text{CV}_2 = \frac{\sigma_2}{\bar{x}_2} \times 100$$

$$= \frac{12}{48} \times 100$$

$$= 25\%$$

$$\begin{aligned}
\mathrm{CV}_3 &= \frac{\sigma_3}{\bar{x}_3}\times 100 \\
&= \frac{2}{12}\times 100 \\
&= 16.67\%
\end{aligned}$$

Since the coefficient of variation of Nayan is least, he is the most consistent.

Example 2

The runs scored by two batsmen A and B in 9 consecutive matches are given below:

A	85	20	62	28	74	5	69	4	13
B	72	4	15	30	59	15	49	27	26

Which of the batsmen is more consistent?

Solution

$$n = 9$$

For the batsman A,

$$\sum x_A = 85+20+62+28+74+5+69+4+13 = 360$$

$$\sum x_A^2 = 85^2+20^2+62^2+28^2+74^2+5^2+69^2+4^2+13^2 = 22700$$

$$\begin{aligned}
\sigma_A &= \sqrt{\frac{\sum x_A^2}{n}-\left(\frac{\sum x_A}{n}\right)^2} \\
&= \sqrt{\frac{22700}{9}-\left(\frac{360}{9}\right)^2} \\
&= 30.37
\end{aligned}$$

$$\bar{x}_A = \frac{\sum x_A}{n} = \frac{360}{9} = 40$$

$$\begin{aligned}
\mathrm{CV}_A &= \frac{\sigma_A}{\bar{x}_A}\times 100 \\
&= \frac{30.37}{40}\times 100 \\
&= 75.925\%
\end{aligned}$$

For the batsman B,

$$\sum x_B = 72+4+15+30+59+15+49+27+26 = 297$$

$$\sum x_B^2 = 72^2+4^2+15^2+30^2+59^2+15^2+49^2+27^2+26^2 = 13837$$

$$\begin{aligned}\sigma_B &= \sqrt{\frac{\sum x_B^2}{n} - \left(\frac{\sum x_B}{n}\right)^2}\\ &= \sqrt{\frac{13837}{9} - \left(\frac{297}{9}\right)^2}\\ &= 21.18\end{aligned}$$

$$\bar{x}_B = \frac{\sum x_B}{n} = \frac{297}{9} = 33$$

$$\begin{aligned}\mathrm{CV}_B &= \frac{\sigma_B}{\bar{x}_B}\times 100\\ &= \frac{21.18}{33}\times 100\\ &= 64.18\%\end{aligned}$$

Since $\mathrm{CV}_B < \mathrm{CV}_A$, the batsman B is more consistent.

Example 3

Two automatic filling machines A and B are used to fill a mixture of cement concrete in a beam. A random sample of beams on each machine showed the following information:

Machine A	32	28	47	63	71	39	10	60	96	14
Machine B	19	31	48	53	67	90	10	62	40	80

Find the standard deviation of each machine and also comment on the performances of the two machines.

[Summer 2015]

Solution

$$n = 10$$

$$\bar{x} = \frac{\sum x}{n} = \frac{460}{10} = 46$$

$$\bar{y} = \frac{\sum y}{n} = \frac{500}{10} = 50$$

Machine A			Machine B		
x	$x-\overline{x}$	$(x-\overline{x})^2$	y	$y-\overline{y}$	$(y-\overline{y})^2$
32	–14	196	19	–31	961
28	–18	324	31	–19	361
47	1	1	48	–2	4
63	17	289	53	3	9
71	25	625	67	17	289
39	–7	49	90	40	1600
10	–36	1296	10	–40	1600
60	14	196	62	12	144
96	50	2500	40	–10	100
14	–32	1024	80	30	900
$\sum x = 460$		$\Sigma(x-\overline{x})^2 = 6500$	$\sum y = 500$		$\Sigma(y-\overline{y})^2 = 5968$

$$\sigma_A = \sqrt{\frac{\sum(x-\overline{x})^2}{n}}$$

$$= \sqrt{\frac{6500}{10}}$$

$$= 25.495$$

$$\sigma_B = \sqrt{\frac{\sum(y-\overline{y})^2}{n}}$$

$$= \sqrt{\frac{5968}{10}}$$

$$= 24.429$$

$$\text{CV}_A = \frac{\sigma_A}{\overline{x}} \times 100$$

$$= \frac{25.495}{46} \times 100$$

$$= 55.423\%$$

$$\text{CV}_B = \frac{\sigma_B}{\overline{y}} \times 100$$

$$= \frac{24.429}{50} \times 100$$

$$= 48.858\%$$

Since $\text{CV}_B < \text{CV}_A$, there is less variability in the performance of the machine B.

EXERCISE 8.5

1. Find the standard deviation of 10 persons whose income in rupees is given below:

312, 292, 227, 235, 269, 255, 333, 348, 321, 299

[**Ans.:** 39.24]

2. Calculate the standard deviation from the following data:

Heights in cm	150	155	160	165	170	175	180
No. of students	15	24	32	33	24	16	6

[**Ans.:** 8.038 cm]

3. Find the standard deviation of the following data:

Size of items	10	11	12	13	14	15	16
Frequency	2	7	11	15	10	4	1

[**Ans.:** 1.342]

4. Calculate the standard deviation for the following frequency distribution:

Class interval	0–4	4–8	8–12	12–16
Frequency	4	8	2	1

[**Ans.:** 3.27]

5. Calculate the standard deviation of the following series:

Marks	0–10	10–20	20–30	30–40	40–50
Frequency	10	8	15	8	4

[**Ans.:** 12.37]

6. Calculate the SD for the following distributions of 300 telephone calls according to their durations in seconds:

Duration (in seconds)	0–30	30–60	60–90	90–120	120–150	150–180	180–210
No. of calls	9	17	43	82	81	44	24

[**Ans.:** 42.51]

7. Calculate the standard deviation from the following data:

Age less than (in years)	10	20	30	40	50	60	70	80
No. of persons	15	30	53	75	100	110	115	125

[**Ans.:** 19.75]

8. Find the standard deviation from the following data:

Mid-value	30	35	40	45	50	55	60	65	70	75	80
Frequency	1	2	4	7	9	13	17	12	7	6	3

[**Ans.:** 11.04]

9. Two cricketers scored the following runs in ten innings. Find who is a better run-getter and who is a more consistent player.

A	42	17	83	59	72	76	64	45	40	32
B	28	70	31	0	59	108	82	14	3	95

[**Ans.:** *A* is a better run-getter and *B* is more consistent.]

10. Two workers on the same job show the following results over a long period of time:

	Worker *A*	Worker *B*
Mean time (in minutes)	30	25
Standard deviation (in minutes)	6	4

[**Ans.:** *B* is more consistent]

8.11 MOMENTS

Moment is the arithmetic mean of the various powers of the deviations of items from their assumed mean or actual mean. If the deviations of the items are taken from the arithmetic mean of the distribution, it is known as *central moment*. If the mean of the first power of deviations are taken, the first moment about the mean is obtained and is denoted by μ_1. The mean of the second power of the deviations gives the second moment about the mean and is denoted by μ_2. Similarly, the mean of the cubes of deviations gives third moment about the mean and is denoted by μ_3. The mean of the fourth power of the deviations from the mean gives the fourth moment about the mean and is denoted by μ_4. Thus, the mean of the r^{th} power of deviations gives the r^{th} moment about mean or r^{th} central moment and is denoted by μ_r.