CHAPTER FIVE

STATISTICS, CORRELATION AND REGRESSION

5.1 INTRODUCTION

In recent decades, the growth of statistic has made itself felt in almost every major phase of human activity, particularly so in the field of Engineering and Science. Everything dealing with the collection, processing, analysis and interpretation of numerical data belongs to the field of statistics. Collection and processing of data is usually referred to as statistical survey. Before any major project work is undertaken, the statistical survey is a must. Only when statistical survey gives green signal, actual start of the work is undertaken. For example, if a Dam is to be constructed on a river, many aspects have to be taken into account. Foremost is the selection of dam site. For making a proper choice, it may be necessary to consider average rainfull in the catchment area for the past say 100 years, the extent of the area which may be submerged, the population which is going to be benefitted, the availability of labour and many other aspects. Good statistical survey should be able to answer all these questions. All such considerations and statistical survey have to be made whenever a new industry is to be started, The success of such projects depends to a great extent upon sound statistical survey. Apart from these basic considerations, modern statistical techniques are widely used in the fields of statistical work, Quality control, reliability needs of the highly complex products of space technology and operation research.

Aim of this work is to introduce to the readers, the simple aspects of collection, classification and enumeration of numerical data, which are so very essential for development of modern statistical techniques, used in engineering fields.

5.2 COLLECTION AND CLASSIFICATION OF DATA

Data collected in a statistical survey as a result of some kind of experimentation is usually large in size and is in the form which is not very useful for arriving at any specific conclusions. The first task is to present this data in a proper form. As a first step, this data which is generally in the form of numerical observations, is arranged either in the ascending or descending order. For example, the set of observations 45, 35, 0, 10, 0, 51, 81, 71, 95, 17, 97, 21, 26, 86, 100, 55, 46, 56, 37, 92 (which are in all 20) is rearranged in ascending order as 0, 0, 10, 17, 21, 26, 35, 37, 45, 51, 55, 56, 71, 81, 86, 92, 95, 97, 100.

This way of presentation immediately reveals that the minimum value of the observation is 0 and maximum is 100. It also indicates that observations are well spread out in the interval (0, 100). In different experiments, these observations could carry different meanings. In some experiments, these figures may indicate the number of syntax errors committed by a group of 20 students in their first attempt to write a computer program. In yet another experiment, these figures may indicate marks obtained out of 100 by a group of 20 students in the paper of numerical computational methods. In an altogether different context, these figures may indicate Rainfall in centimeters in a certain catchment area for the past 20 years. For development of statistical techniques it is unimportant, what is exactly represented by these observations. In presentation of data, these observations are represented by symbol x, called in statistical language, a variate (variable).

After arranging the data in ascending or descending order, to make it more compact, it is presented in a tabular form consisting of columns headed by symbols x and f. The column headed by x consists of various observations recorded out of experimentation, arranged in proper order, and column headed by f contains entries which indicate number of times particular value of x occurs.

Consider the Table 5.1, which shows various values of x and f. It shows that the value of $x = 1$ is recorded twice, $x = 4$ occurs six times, $x = 8$ occurs four times etc.

Table 5.1

x	f
1	2
2	3
3	5
4	6
5	10
6	6
7	4
8	4
9	3
10	2
–	$\Sigma f = 45$

The total numbers of observations being $\Sigma f = 45$. In statistical language, this table means $x = 1$ has frequency 2, $x = 4$ has frequency 6 and so on. This way of arrangement of data is called *frequency distribution.* In the above example, the range of variate is from $x = 1$ to $x = 10$. When the range is wide and the total number of observations is very large, the data can be expressed in still more compact form by dividing the range in class intervals.

Consider the table given on next page (Table 5.2). Here the range of variate (0, 100) is divided into 10 class intervals each of width 10. The class interval 0 – 10 has width 10, the lower limit 0 and the upper limit 10.

$\frac{10+0}{2} = 5$ is the middle value of the class interval and 16 is the frequency corresponding to this class interval. The middle value $x = 5$ represents the class interval (0 – 10) of $f = 16$ is taken as frequency of variate x. This way of representing the data is

called *Grouped frequency distribution*. In such type of presentation, the class intervals must be well defined. One such way of defining the class interval is that, all the values of $x = 0$ and above but less than 10 are included in the class interval 0 – 10. The total frequency of all such observations is 16 and is the frequency of class interval 0 – 10 or is the frequency of variate $x = 5$.

Table 5.2

C.I. (Class interval)	Mid-value x	Frequency f
0 – 10	5	16
10 – 20	15	18
20 – 30	25	20
30 – 40	35	22
40 – 50	45	40
50 – 60	55	45
60 – 70	65	35
70 – 80	75	20
80 – 90	85	19
90 – 100	95	15
Total	–	$\Sigma f = 250$

Similarly all the observations having the value $x = 10$ and above but less than 20 are included in the class interval 10 – 20 and so on. Slight change in the definition of last class interval is made. Here all the values of $x = 90$ and above and less than or equal to 100 are included in the class interval 90 – 100. $\Sigma f = 250$ gives the total frequency which is sometimes denoted by N.

In presenting the data in Grouped frequency distribution form, the following points must be noted :

(i) The class interval must be well defined that is there must not be any ambiguity about the inclusion of value of x in one or the other class interval. In the Table 5.2, the way of defining class interval enables us to put $x = 10$ in the class interval 10 – 20 while $x = 100$ is put in the interval 90 – 100.

(ii) The class intervals must be exhaustive that is no observation should escape classification. For this, the entire range of observations should be divided into well defined class intervals.

(iii) The width of the class interval should be uniform as far as possible.

(iv) The number of class intervals should neither be too large nor too small. Depending upon the range of variate x and the total frequency of observations, the total number of class intervals is divided into about 10 to 25 class intervals.

Sometimes the additional column of cumulative frequency (c.f.) suppliments the grouped frequency distribution or frequency distribution table.

In the Table 5.3, the number 76 against x = 35 shows the total frequency upto and including the observation x = 35 which is the middle value of the interval (30 – 40).

Table 5.3

C.I.	Mid-value x	Frequency f	Cumulative frequency c.f.
0 – 10	5	16	16
10 – 20	15	18	34
20 – 30	25	20	54
30 – 40	35	22	76
40 – 50	45	40	116
50 – 60	55	45	161
60 – 70	65	35	196
70 – 80	75	20	216
80 – 90	85	19	235
90 – 100	95	15	250
Total	–	**$\Sigma f = 250$**	**N = 250**

Graphical Representation of Data

To observe the data at a glance, it is exhibited by following graphical methods :

1. Histogram : A Histogram is drawn by constructing rectangles over the class intervals, such that the areas of rectangles are proportional to the class frequencies.

If the class intervals are of equal width, the heights of the rectangles will be proportional to the class frequencies themselves, otherwise these would be proportional to the ratios of the frequencies to the width of the classes (See Fig. 5.1).

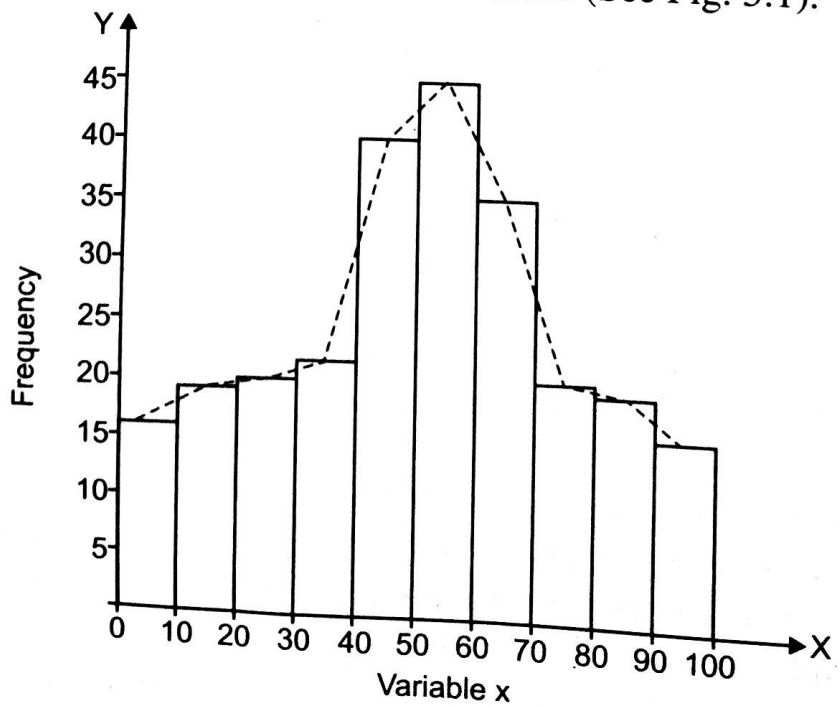

Fig. 5.1

2. Frequency Polygon : Consider the set of points (x, f), where x is the middle value of the class interval and f is the corresponding frequency. If these set of points are joined by straight lines, they form a frequency polygon. It is shown by dotted lines in Fig. 5.1.

3. Cumulative Frequency Curve or The Ogive : Taking upper limit of classes of x co-ordinate and corresponding cumulative frequency as y co-ordinate, if the points are plotted and then joined by free hand curve, it gives what is called as ogive (See Fig. 5.2).

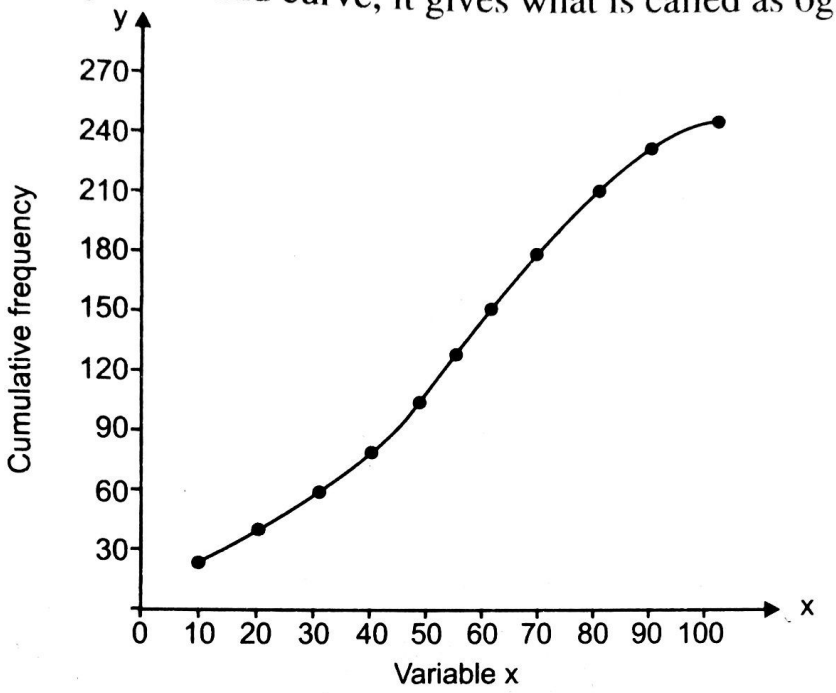

Fig. 5.2

5.3 LOCATION OF CENTRAL TENDENCY

After collecting the data and arranging it in the proper order in the form of frequency distribution or grouped frequency distribution, next task is to study this data carefully and to draw valid conclusions. If data collected relates to marks obtained by the students in Mathematics paper, it should be able to reveal the general performance of the students. Whether the class contains large number of good students or the overall calibre of students is medicore, all this must be inferred from the data. If the numerical data collected relates to the industrial project, the whole success of the project will depend upon the appropriate conclusions drawn from the study of this data. The first step in this direction is the location of central tendency. It means what is represented by data by and large. Whether the data is favourable to a particular project or not will depend upon the criterion that is decided upon. But overall picture must be exhibited. This overall picture or central tendency of the data is known by obtaining what we call the Mean or Average. There are various methods to calculate the mean or the average. Depending upon the project under study, the particular method is selected. Various types of measures of central tendency are as given below :

(1) Arithmetic mean (2) Geometric mean
(3) Harmonic mean (4) Median (5) Mode.

Out of these, Arithmetic mean is of greater importance and serves the purpose in many cases. Now, we see how these measures are calculated.

5.3.1 Arithmetic Mean

Consider the variate x which takes n values $x_1, x_2, x_3 \ldots\ldots x_n$, then the Arithmetic mean (A.M.) is denoted by $\bar{x}$ and is given by,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n}$$

If the data is presented in the form of frequency distribution

x	x_1	x_2	x_3	.....	x_n
f	f_1	f_2	f_3	.....	f_n

then arithmetic mean $\bar{x}$ is given by

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

$$= \frac{\sum fx}{N}$$

where $N = f_1 + f_2 + \ldots\ldots + f_n$ is the total frequency

ILLUSTRATIONS

Ex. 1 : *Find the Arithmetic mean for the following distribution :*

x	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
f	*4*	*5*	*12*	*12*	*13*	*16*	*15*	*13*	*12*	*5*	*6*

Sol. : Writing the tabulated values as :

x	f	x × f
0	4	0
1	5	5
2	12	24
3	12	36
4	13	52
5	16	80
6	15	90
7	13	91
8	12	96
9	5	45
10	6	60
Total	**$\Sigma f = 113$**	**$\Sigma fx = 579$**

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{579}{113} = 5.12 \text{ (approximately)}$$

To reduce the calculations, we consider the variable $d = x - A$.

Where, A is middle value or value near to it in the range of variable x, A is sometimes called assumed mean.

Now we can write

$$f \times d = f \times x - f \times A$$

Or $$\Sigma fd = \Sigma fx - \Sigma fA$$

Dividing by Σf throughout

$$\frac{\Sigma fd}{\Sigma f} = \frac{\Sigma fx}{\Sigma f} - \frac{\Sigma fA}{\Sigma f}$$

i.e. $$\frac{\Sigma fd}{\Sigma f} = \bar{x} - \frac{A \Sigma f}{\Sigma f}$$

(A being constant is taken outside the Σ notation)

$$\frac{\Sigma fd}{\Sigma f} = \bar{x} - A$$

Or $$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} = A + \bar{d}$$ [$\bar{d}$ is the mean of the variable d]

fd and Σ fd are smaller numbers as compared to fx and Σ fx, which result in the reduction of the calculations.

Further reduction in calculations can be achieved by taking

$$u = \frac{x - A}{h} \text{ or } \frac{d}{h}$$

that gives $$hu = x - A$$

Then proceeding as before, we get

$$h \frac{\Sigma fu}{\Sigma f} = \bar{x} - A$$

Or $$\bar{x} = A + h \frac{\Sigma fu}{\Sigma f}$$

This formula is mostly used in grouped frequency distribution, where, h is chosen to be equal to the width of the class interval.

Ex. 2 : *Marks obtained in a paper of statistics are given in the following table.*

Marks obtained	***No. of students***
0 – 10	*8*
10 – 20	*20*
20 – 30	*14*
30 – 40	*16*
40 – 50	*20*
50 – 60	*25*
60 – 70	*13*
70 – 80	*10*
80 – 90	*5*
90 – 100	*2*

Find the Arithmetic mean of the distribution.

Sol. : Preparing the table as : A = 45, h = 10.

C.I.	Mid-value x	f	$u = \frac{x-45}{10}$	f × u
0 – 10	5	8	– 4	– 32
10 – 20	15	20	– 3	– 60
20 – 30	25	14	– 2	– 28
30 – 40	35	16	– 1	– 16
40 – 50	45	20	0	0
50 – 60	55	25	1	25
60 – 70	65	13	2	26
70 – 80	75	10	3	30
80 – 90	85	5	4	20
90 – 100	95	2	5	10
Total	–	**$\sum f = 133$**	–	**$\sum fu = -25$**

$$\bar{x} = A + h\frac{\sum fu}{\sum f} = 45 + 10\left(\frac{-25}{133}\right)$$

$$= 45 + 10\left(\frac{-25}{133}\right) = 45 - \frac{250}{133} = 43.12$$

Joint Arithmetic Mean (Mean of composite series)

Consider two sets of data

1. $x_1, x_2, \ldots, x_{n_1}$ containing n_1 items
2. $y_1, y_2, \ldots, y_{n_2}$ containing n_2 items

$\therefore$ $\bar{x}$, the mean of first set is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n_1}$$

$\therefore$ $$n_1 \bar{x} = x_1 + x_2 + \ldots + x_{n_1}$$

and the mean of second set is given by

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \ldots + y_{n_2}}{n_2}$$

$\therefore$ $$n_2 \bar{y} = y_1 + y_2 + \ldots + y_{n_2}$$

Hence, by definition the joint arithmetic mean $\bar{z}$ is given by

$$\bar{z} = \frac{(x_1 + x_2 + x_3 + \ldots + x_{n_1}) + (y_1 + y_2 + y_3 + \ldots + y_{n_2})}{n_1 + n_2}$$

$\therefore$ $$\bar{z} = \frac{n_1 \bar{x} + n_2 \bar{y}}{n_1 + n_2} \quad \ldots (A)$$

(A) gives joint Arithmetic Mean (A.M.) of the composite series.

Same type of formula holds good for sets of data presented in frequency distribution form. Consider two sets of data :

Set 1	
x	**f**
x_1	f_1
x_2	f_2
x_3	f_3
...	...
...	...
x_{n_1}	f_{n_1}
	$\Sigma f = N_1$

Set 2	
y	**f**
y_1	f_1
y_2	f_2
y_3	f_3
...	...
...	...
y_{n_1}	f_{n_1}
	$\Sigma f = N_2$

Means $\bar{x}$, $\bar{y}$ for two sets are given by

$$\bar{x} = \frac{\Sigma fx}{N_1} \qquad N_1 \bar{x} = \Sigma fx$$

$$\bar{y} = \frac{\Sigma fy}{N_2} \qquad N_2 \bar{y} = \Sigma fy$$

Hence $\bar{z}$, the joint mean given by

$$\bar{z} = \frac{\Sigma fx + \Sigma fy}{N_1 + N_2} = \frac{N_1 \bar{x} + N_2 \bar{y}}{N_1 + N_2} \qquad \ldots (A)$$

Ex. 3 : *Marks obtained in paper of Applied Mechanics by a group of Computer and Electronics students are as given in following tables :*

Group (A) of Computer students :

Marks obtained	*No. of students*
0 – 10	*5*
10 – 20	*6*
20 – 30	*15*
30 – 40	*15*
40 – 50	*9*
	$\Sigma f = 50$

Group (B) of Electronics students :

Marks obtained	*No. of students*
0 – 10	*8*
10 – 20	*15*
20 – 30	*18*
30 – 40	*13*
40 – 50	*6*
	$\Sigma f = 60$

Find the Joint mean of the two groups.

Sol. : For group (A) :

C.I.	Mid-value x	f	f × x
0 – 10	5	5	25
10 – 20	15	6	90
20 – 30	25	15	375
30 – 40	35	15	525
40 – 50	45	9	405
Total	–	$N_1 = \sum f = 50$	$\sum fx = 1420$

For group (B) :

C.I.	Mid-value x	f	f × x
0 – 10	5	8	40
10 – 20	15	15	225
20 – 30	25	18	450
30 – 40	35	13	455
40 – 50	45	6	270
Total	–	$N_2 = \sum f = 60$	$\sum fy = 1440$

Mean $\bar{x}$ of group (A) is given by,

$$\bar{x} = \frac{\sum fx}{\sum f} \Rightarrow \bar{x} \sum f = N_1 \bar{x} = \sum fx = 1420$$

Mean $\bar{y}$ of group (B) is given by,

$$\bar{y} = \frac{\sum fy}{\sum f} \Rightarrow \bar{y} \sum f = N_2 \bar{y} = \sum fy = 1440$$

Common mean $\bar{z}$ is given by,

$$\bar{z} = \frac{N_1 \bar{x} + N_2 \bar{y}}{N_1 + N_2} = \frac{1420 + 1440}{50 + 60}$$

$$= \frac{2860}{110} = 26$$

Ex. 4 : *Calculate arithmetic mean for the following frequency distribution :*

Observations (x)	*103*	*110*	*112*	*118*	*95*
Frequency (f)	*4*	*6*	*10*	*12*	*3*

Solution : We solve the problem by both the methods.

1. Direct Method

x	f	fx
103	4	$103 \times 4 = 412$
110	6	$110 \times 6 = 660$
112	10	$112 \times 10 = 1120$
118	12	$118 \times 12 = 1416$
95	3	$95 \times 3 = 285$
Total	**N = 35**	$\sum f_i x_i = 3893$

$$\therefore \quad \bar{x} = \frac{\sum fx}{\sum f} = \frac{3893}{35} = 111.2286$$

Ex. 5 : *Arithmetic mean of weight of 100 boys is 50 kg and the arithmetic mean of 50 girls is 45 kg. Calculate the arithmetic mean of combined group of boys and girls.*

Solution : Let $\bar{X}_1$ and N_1 be the mean and size of group of boys and $\bar{Y}$ and N_2 be the mean and size of group of girls. So $N_1 = 100$, $\bar{X} = 50$, $N_2 = 50$, $\bar{Y} = 45$. Hence, combined mean is

$$Z = \frac{N_1\bar{X} + N_2\bar{Y}}{N_1 + N_2} = \frac{(100 \times 50) + (50 \times 45)}{100 + 50} = \frac{7250}{150} = 48.3333$$

Ex. 6 : *The mean weekly salary paid to 300 employees of a firm is ₹ 1,470. There are 200 male employees and the remaining are females. If mean salary of males is ₹ 1,505, obtain the mean salary of females.*

Solution : Suppose $\bar{X}$ and N_1 are mean and group size of males. $\bar{Y}$ and N_2 are mean and size of group of females, $\bar{x}_c$ mean of all the employees considered together.

Now, $$Z = \frac{N_1\bar{X} + N_2\bar{Y}}{N_1 + N_2}$$

$$\therefore \quad 1470 = \frac{(200 \times 1505) + (100 \times \bar{Y})}{200 + 100}$$

$$\therefore \quad 1470 = \frac{301000 + 100\bar{Y}}{300}$$

$$\therefore \quad 441000 = 301000 + 100Y$$

$$\therefore \quad 4410 = 3010 + Y$$

$$\therefore \quad \bar{Y} = 1,400 ₹$$

5.3.4 Median

Median of a distribution is the value of the variable (or variate) which divides it into two equal parts. It is the value such that the number of observations above it is equal to the number of observations below it. Sometimes, Median is called positional average.

In case of ungrouped data, if the number of observations is odd, then the median is the middle value of the set of observations after they are arranged in ascending or descending order. For even number of observations, it is the arithmetic mean of the two middle terms. Thus for the observations

$x = 1, 5, 9, 11, 21, 24, 27, 30$, the middle terms are 11 and 21 and median $= \frac{11 + 21}{2} = 16.$

For a data presented in the form of frequency distribution :

x	x_1	x_2	.....	x_n
f	f_1	f_2	.....	f_n

$\Sigma f = N$

We prepare the cumulative frequency column. Then consider cumulative frequency (c.f.) equal to $\frac{N}{2}$ or just greater than $\frac{N}{2}$, the corresponding value of x is the median.

ILLUSTRATIONS

Ex. 1 : *Obtain the median of the distribution :*

x	*1*	*3*	*5*	*7*	*9*	*11*	*13*	*15*	*17*
f	*3*	*6*	*8*	*12*	*16*	*16*	*15*	*10*	*5*

Sol. : Preparing the table as :

x	f	c.f.
1	3	3
3	6	9
5	8	17
7	12	29
9	16	45
11	16	61
13	15	76
15	10	86
17	5	91
Total	**$\Sigma f = 91$**	**–**

Here the total frequency N = 91; $\frac{N}{2} = 45.5$.

The value of c.f. just greater than 45.5 is 61, the corresponding value of x is 11 and thus median is 11. In case of grouped frequency distribution, the class corresponding to the c.f. just greater than $\frac{N}{2}$ is called the median class and the value of median is obtained by the formula :

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

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 c.f. of the class preceding the median class

Ex. 2 : *Wages earned in Rupees per day by the labourers are given by the table :*

Wages in ₹	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*
No. of labourers	*5*	*8*	*13*	*10*	*8*

Find the median of the distribution.

Sol. :

Wages in ₹ C.I.	**No. of labourers f**	**(c.f.)**
10 – 20	5	5
20 – 30	8	13
30 – 40	13	26
40 – 50	10	36
50 – 60	8	44
Total	**$\Sigma f = N = 44$**	–

Here $\frac{N}{2} = \frac{44}{2} = 22$

Cumulative frequency (c.f.) just greater than 22 is 26 and the corresponding class is 30 – 40.

Using formula to calculate median,

$l = 30$, f = 13, h = 10, $\frac{N}{2} = 22$, c = 13

$$\text{Median} = 30 + \frac{10}{13}(22 - 13)$$

$$= 30 + \frac{10}{13}(9) = 30 + \frac{90}{13} = 36.923$$

5.3.5 Mode

It is the value of the variate which occurs most frequently in a set of observations, or is the value of variate corresponding to maximum frequency.

In case of grouped frequency distribution, Mode is given by the formula :

$$\text{Mode} = l + \frac{h\,(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)}$$

$$= l + \frac{h\,(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

Here,

l is the lower limit of the modal class

h is the width of the modal class

f_1 is the frequency of the modal class

f_0 is the frequency of the class preceding to the modal class

f_2 is the frequency of the class succeeding to the modal class.

ILLUSTRATION

Ex. 9 : *Find the Mode for the following distribution :*

C.I.	*0 – 10*	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*	*60 – 70*
f	*4*	*7*	*8*	*12*	*25*	*18*	*10*

Sol. : Here C.I. 40 – 50 corresponding to which f = 25 is maximum, is the modal class.

$l = 40$, h = 10, $f_1 = 25$, $f_0 = 12$, $f_2 = 18$

$$\text{Mode} = 40 + \frac{10\,(25 - 12)}{(2 \times 25 - 12 - 18)}$$

$$= 40 + \frac{130}{20} = 40 + 6.5 = 46.5$$

So far we have considered various ways in which average can be calculated. It is clear that no single average is suitable for all types of data. Arithmetic mean, Geometric mean and Harmonic mean are rigidly defined and are based on all the observations, they are suitable for further mathematical treatment. They are not much affected by fluctuations of sampling. In fact among all the averages, Arithmetic mean is least affected by fluctuations. Geometric mean becomes zero if any one of the observations is zero. Geometric and Harmonic means are not easy to understand and are difficult to compute. They give greater importance to small items and are useful when small items have to be given a very high weightage. Median and Mode are not amenable to algebraic treatment. Their main advantage is that they are not affected by extreme values, but compared to Arithmetic mean they are affected much by fluctuations of sampling. All the averages have merits and demerits, but Arithmetic mean because of its simplicity and its stability is much more familiar to a lyman. It has wide applications in statistical theory and is considered as best among all the averages.

5.4 DISPERSION

After calculation of the average using any of the five methods discussed in previous section, question arises whether the average calculated gives correct information about the central tendency of the data, the purpose for which it is calculated. Main point to be discussed is whether the average is true representative of the data or not. As an illustration, consider the two sets of observations :

(i) 5, 10, 15, 20, 25.

(ii) 13, 14, 15, 16, 17.

The Arithmetic mean of both these sets is 15. It is obvious that 15 is better average for second than the first, because the observations in the second set are much closer to the value 15 as compared to the first set. In the second set, the values of the variate are much less scattered or dispersed from the mean as compared to the first. There are two widely accepted ways of measuring the degree of scatteredness from the mean. These are :

(i) Mean deviation

(ii) Standard deviation.

These are the measures of dispersion, which decide whether the average truely represents the given data or not. Besides these two standard measures, there are other measures such as Range and Quartile deviation or semi-interquartile range. But these are not as much of consequence. We shall now discuss about the two measures of dispersion mentioned earlier.

(i) Mean Deviation : For a frequency distribution

x	x_1	x_2	.....	x_n
f	f_1	f_2	.....	f_n

mean deviation from the average A (usually Arithmetic mean or at most median or mode) is given by,

$$\text{Mean deviation} = \frac{1}{N} \Sigma f \, |x - A|$$

where, $N = \Sigma f$ is the total frequency, $|x - A|$ represents the modulus or the absolute value of the deviation $(x - A)$ ignoring the – ve sign. It can be broadly stated that when deviation is a small number, the average is good.

ILLUSTRATION

Ex. 1 : *Calculate Arithmetic mean and Mean deviation of the following frequency distribution :*

x	*1*	*2*	*3*	*4*	*5*	*6*
f	*3*	*4*	*8*	*6*	*4*	*2*

Sol. : Preparing the table :

x	f	x × f	x – A	|x – A|	f × | x – A |
1	3	3	– 2.37	2.37	7.11
2	4	8	– 1.37	1.37	5.48
3	8	24	– 0.37	0.37	2.96
4	6	24	0.63	0.63	3.78
5	4	20	1.63	1.63	6.52
6	2	12	2.63	2.63	5.26
Total	**$\Sigma f = 27$**	**$\Sigma f x = 91$**	–	–	**$\Sigma f \times \|x - A\| = 31.11$**

$$\text{A.M.} = A = \frac{\Sigma fx}{\Sigma f} = \frac{91}{27} = 3.37 \quad \text{(approximately)}$$

$$\text{Mean deviation} = \frac{\Sigma f \times |x - A|}{\Sigma f} = \frac{31.11}{27} = 1.152 \quad \text{(approximately).}$$

(ii) Standard Deviation : *It is defined as 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 symbol σ.

For a frequency distribution (x, f),

$$\sigma = \sqrt{\frac{1}{N} \Sigma f \left(x - \bar{x}\right)^2}$$

where, $\bar{x}$ is A.M. of the distribution and $N = \Sigma f$.

The square of the standard deviation is called variance, denoted by V.

Thus, $$V = \sigma^2 = \frac{1}{N} \Sigma f \cdot \left(x - \bar{x}\right)^2$$

The step of squaring the deviations $\left(x - \bar{x}\right)$ overcomes the drawback of ignoring the signs in Mean deviation. Standard deviation is also suitable for further mathematical treatment. Moreover among all the measures of dispersion, standard deviation is affected by least fluctuations of sampling, hence it is considered as most reliable measure of dispersion.

Root mean square deviation is given by

$$S = \sqrt{\frac{1}{N} \Sigma f \, (x - A)^2}$$

where, A is any arbitrary number.

S^2 is called Mean square deviation. When $A = \bar{x}$, the Arithmetic mean, Root mean square deviation becomes equal to the standard deviation.

(iii) Relation Between σ and S : By definition, we have

$$S^2 = \frac{1}{N}\sum f(x-A)^2$$

$$= \frac{1}{N}\sum f\left(x-\bar{x}+\bar{x}-A\right)^2$$

$$= \frac{1}{N}\sum f\left[\left(x-\bar{x}\right)^2 + 2\left(x-\bar{x}\right)\left(\bar{x}-A\right) + \left(\bar{x}-A\right)^2\right]$$

$$= \frac{1}{N}\sum f\left(x-\bar{x}\right)^2 + 2\left(\bar{x}-A\right)\frac{1}{N}\sum f\left(x-\bar{x}\right) + \left(\bar{x}-A\right)^2\frac{\sum f}{N}$$

Note that $\left(\bar{x}-A\right)$ being constant, is taken outside the summation.

Now since $\frac{1}{N}\sum f\left(x-\bar{x}\right) = \frac{1}{N}\sum fx - \bar{x}\cdot\frac{1}{N}\sum f = \bar{x}-\bar{x} = 0$

$$\therefore \qquad S^2 = \frac{1}{N}\sum f\left(x-\bar{x}\right)^2 + \left(\bar{x}-A\right)^2, \quad \text{as } \sum f = N$$

Thus, $\qquad S^2 = \sigma^2 + d^2,\ d = \bar{x} - A$

If $\bar{x} = A$, thus S^2 would be least as $d = 0$.

Thus Mean square deviation (S^2) and consequently Root mean square (S) deviation are least when deviations are taken from $A = \bar{x}$.

(iv) Method of Calculating σ :

$$\sigma^2 = \frac{1}{N}\sum f\left(x-\bar{x}\right)^2$$

$$= \frac{1}{N}\sum f\left(x^2 - 2x\bar{x} + \bar{x}^2\right)$$

$$= \frac{1}{N}\sum fx^2 - \frac{2\bar{x}}{N}\sum fx + \bar{x}^2\cdot\frac{\sum f}{N}$$

$$= \frac{1}{N}\sum fx^2 - 2\bar{x}^2 + \bar{x}^2 \left[\frac{\sum fx}{N} = \bar{x},\ \frac{\sum f}{N} = 1\right]$$

$$= \frac{1}{N}\sum fx^2 - \bar{x}^2$$

$$= \frac{1}{N}\sum fx^2 - \left(\frac{1}{N}\sum fx\right)^2$$

Usually, product terms fx and fx^2 are large, hence to reduce the volume of calculations, we proceed as follows :

$$\sigma^2 = \frac{1}{N}\sum f\left(x-\bar{x}\right)^2$$

$$= \frac{1}{N}\sum f\left(x-A+A-\bar{x}\right)^2, \qquad \text{(where A is arbitrary number)}$$

$$= \frac{1}{N}\sum f\left[(x-A)^2 + 2(x-A)\left(A-\bar{x}\right) + \left(A-\bar{x}\right)^2\right]$$

$$= \frac{1}{N}\sum f(x-A)^2 + \frac{2}{N}\sum\left(A-\bar{x}\right)\sum f(x-A) + \left(A-\bar{x}\right)^2\frac{\sum f}{N}$$

Let $d = x - A$ then using

$$\bar{x} = A + \frac{1}{N}\sum fd$$

$$\sigma^2 = \frac{1}{N}\sum fd^2 + \frac{2}{N}\left[A - A - \frac{1}{N}\sum fd\right]\sum fd + \left[A - A - \frac{1}{N}\sum fd\right]^2 \cdot 1$$

$$= \frac{1}{N}\sum fd^2 - \frac{2}{N^2}\left(\sum fd\right)^2 + \frac{1}{N^2}\left(\sum fd\right)^2$$

$$= \frac{1}{N}\sum fd^2 - \frac{1}{N^2}\left(\sum fd\right)^2$$

$$= \frac{1}{N}\sum fd^2 - \left(\frac{\sum fd}{N}\right)^2$$

Or $$\sigma = \sqrt{\frac{1}{N}\sum f d^2 - \left(\frac{\sum fd}{N}\right)^2} \qquad \ldots (A)$$

Terms fd, fd^2 are numerically smaller as compared to fx, fx^2 and use of formula (A) reduces the calculations considerably in obtaining σ.

To reduce the calculations further, and in dealing with data presented in grouped frequency distribution form, we put $u = \frac{x-A}{h}$, where h is generally taken as width of class interval

Thus $u = \frac{d}{h}$ or $d = hu$ putting $d = hu$ in formula (A)

$$\sigma = \sqrt{\frac{1}{N}\sum f h^2 u^2 - \left(\frac{\sum f hu}{N}\right)^2}$$

$$= h\sqrt{\frac{1}{N}\sum f u^2 - \left(\frac{\sum f u}{N}\right)^2} \qquad \ldots (B)$$

Formula (B) is quite useful for data presented in grouped frequency distribution form.

Ex. 1 : *Calculate standard deviation for the following frequency distribution. Decide whether A.M. is good average.*

Wages in Rupees earned per day	*0 – 10*	*10 – 20*	*20 – 30*	*30 – 40*	*40 – 50*	*50 – 60*
No. of labourers	*5*	*9*	*15*	*12*	*10*	*3*

Sol. : Preparing the table for the purpose of calculations.

Wages earned C.I.	**Mid-value x**	**Frequency f**	$\mathbf{u = \frac{x - 25}{10}}$	**fu**	**fu^2**
0 – 10	5	5	– 2	– 10	20
10 – 20	15	9	– 1	– 9	9
20 – 30	25	15	0	0	0
30 – 40	35	12	1	12	12
40 – 50	45	10	2	20	40
50 – 60	55	3	3	9	27
Total	**–**	**$\sum f = 54$**	**–**	**$\sum fu = 22$**	**$\sum fu^2 = 108$**

Using formula (B),

$$\sigma = 10\sqrt{\frac{1}{54} \times 108 - \left(\frac{22}{54}\right)^2}$$

$$= 10\sqrt{2 - 0.166} = 13.54 \text{ approximately}$$

In this problem,

$$\text{A.M.} = 25 + h\,\frac{\sum fu}{N} = 25 + 10\ (0.4074) = 29.074$$

$\sigma = 13.54$ is quite a large value and Arithmetic mean 29.074 is not a good average.

which is the required result.

Ex. 4 : *Fluctuations in the Aggregate of marks obtained by two groups of students are given below. Find out which of the two shows greater variability.*

Group A	518	519	530	530	544	542	518	550	527	527	531	550	550	529	528
Group B	825	830	830	819	814	814	844	842	842	826	832	835	835	840	840

(Dec. 2012)

Sol. : To solve this problem, we have to determine coefficient of variation $\frac{\sigma}{A.M.} \times 100$ in each case. First we present the data in frequency distribution form.

For Group A :

x	f	d = x – 530	d^2	fd	fd^2
518	2	– 12	144	– 24	288
519	1	– 11	121	– 11	121
527	2	– 3	9	– 6	18
528	1	– 2	4	– 2	4
529	1	– 1	1	– 1	1
530	2	0	0	0	0
531	1	1	1	1	1
542	1	12	144	12	144
544	1	14	196	14	196
550	3	20	400	60	1200
Total	**$\Sigma f = 15$**	–	–	**$\Sigma fd = 43$**	**1973**

$$\text{A.M.} = 530 + \frac{\Sigma f d}{\Sigma f}$$

$$= 530 + \frac{43}{15} = 532.866$$

$$\sigma = \sqrt{\frac{1}{N} \Sigma f d^2 - \left(\frac{\Sigma fd}{N}\right)^2}$$

$$= \sqrt{\frac{1973}{15} - \left(\frac{43}{15}\right)^2}$$

$$= \sqrt{131.533 - 8.218}$$

$$= 11.105$$

$$\text{Coefficient of variation} = \frac{\sigma}{\text{A.M.}} \times 100$$

$$= \frac{11.105}{532.866} \times 100 = 2.0840$$

For Group B :

x	f	d = x − 830	d^2	fd	fd^2
814	2	− 16	256	− 32	512
819	1	− 11	121	− 11	121
825	1	− 5	25	− 5	25
826	1	− 4	16	− 4	16
830	2	0	0	0	0
832	1	2	4	2	4
835	2	5	25	10	50
840	2	10	100	20	200
842	2	12	144	24	288
844	1	14	196	14	196
Total	**$\sum f = 15$**	−	−	**$\sum fd = 18$**	**$\sum fd^2 = 1412$**

$$\text{A.M.} = 830 + \frac{18}{15} = 831.2$$

$$\sigma = \sqrt{\frac{1412}{15} - \left(\frac{18}{15}\right)^2} = \sqrt{94.133 - 1.44} = 9.628$$

$$\text{Coefficient of variation} = \frac{9.628}{831.2} \times 100 = 1.158$$

Coefficient of variation of group A is greater than that of group B.

∴ Group A has greater variability, or Group B is more consistent.

1. *Find the Arithmetic Mean, Median and Standard deviation for the following frequency distribution.*

x	5	9	12	15	20	24	30	35	42	49
f	3	6	8	8	9	10	8	7	6	2

Ans. $\bar{x} = 22.9851$, $M = 20$, $\sigma = 11.3538$

2. *Following table gives the Marks obtained in a paper of statistics out of 50, by the students of two divisions :*

C.I.	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
Div. A f	2	6	8	8	15	18	12	11	9	4
Div. B f	3	5	7	9	12	16	11	5	6	2

Find out which of the two divisions show greater variability.

Also find the common mean and standard deviation.

Ans. B has greater variability

$\bar{x} = 26.1458$

$\sigma = 11.1267$

4. The Mean and Standard deviation of 25 items is found to be 11 and 3 respectively. It was observed that one item 9 was incorrect. Calculate the Mean and Standard deviation if :

(i) the wrong item is omitted.

(ii) it is replaced by 13.

Ans. (i) $\bar{x} = 11.08, \sigma = 3.345$

(ii) $\bar{x} = 11.16, \sigma = 2.9915$

5. Age distribution of 150 life insurance policy-holders is as follows :

Age as on nearest birthday	Number
15 – 19.5	10
20 – 24.5	20
25 – 29.5	14
30 – 34.5	30
35 – 39.5	32
40 – 44.5	14
45 – 49.5	15
50 – 54.5	10
55 – 59.5	5

Calculate mean deviation from median age.

Ans. M.D. = 8.4284