Chapter 8

Data and Their Representations

8.1 Introduction

Statistics is the science which deals with the collection, organisation, analysis and interpretation of the numerical data.

Collection and analysis of numerical data is necessary in studying many problems such as the problem of economic development of the country, educational development, the problem of health and population, the problem of agricultural development etc.

In this chapter, we shall study this branch of mathematics with collection, classification, presentation and analysis of data. We shall learn how to classify the given data into ungrouped and grouped frequency distributions.

8.2 Collection of Data

In any field of investigation, the first step is to collect the relavant data. and to analyse it.

Data are said to be **primary** if the investigator himself is responsible for the collection of data. Such as voters'lists, data collected in census-questionnaire etc.

It is not always possible for an investigator to collect data due to many reasons. In that case, he/she may use data collected by other agency in the form of published reports. They are called **secondary data.** Data may be primary for one individual or agency but it becomes secondary for other using the same data.

8.3 Presentation of Data

After the collection of data the next step to the investigator is to find ways to organise them in order to study their important features. Such an arrangement of data is called **presentation of data.**

Suppose there are 20 students in a class. The marks obtained by them in a mathematics test (out of 100) are as follows:

36, 59, 65, 56, 88, 27, 56, 72, 65, 74
45, 56, 61, 56, 31, 33, 72, 61, 76, 56,

The data in this form is called raw data. Each entry is called a **value or observation.**

We arrange these numbers in ascending order:

27, 31, 33, 36, 45, 56, 56, 56, 56, 56,
59, 61, 61, 65, 65, 72, 72, 74, 76, 89 ...(1)

Now you can get the following information:

(a) Highest marks obtained : 89
(b) Lowest marks obtained : 27
(c) Number of students who got 56 marks: 5
(d) Number of students who got marks more than 60 : 9

The data arranged in the above form, are called **arrayed data.** Presentation of data in this form is time cousuming, when the number of observations is large.

To make the data more informative we can present these in a tabular form as shown below:

Marks in Mathematics of 20 students

Marks	Number of Students
27	1
31	1
33	1
36	1
45	1
56	5
59	1
61	2
65	2
70	2
74	1
76	1
89	1
Total	20

Such a table is called a **frequency distribution table for ungrouped data or ungrouped frequency table.**

Note: When the number of observations is large, it may be tideous to find the frequencies by simple counting. In such cases, we make use of bars (|), called **tally marks**).

In order to get condensed form of the data (when the number of observation is large), we classify the data into **classes** or **groups** as below:

Frequency Table of the marks obtained by 20 students in a mathematics test

Class Interval (Marks out of 100)	Tally Marks	Frequency
27-33	|||	3
34-40	|	1
41-47	|	1
48-54	–	0
55-61	~~||||~~ |||	8
62-68	||	2
69-75	|||	3
76-82	|	1
83-89	|	1
Total		**20**

The above table is called a **frequency distribution table** for grouped data.

Now let us consider the following frequency distribution table which gives the weight of 50 students of a class:

Weight (in kg)	Number of Students
31-35	10
36-40	7
41-45	15
45-50	4
51-55	2
56-60	3
61-65	4
66-70	3
71-75	2
Total	**50**

Suppose two students of weights 35.5 kg and 50.54 kg are admitted in this class.

In which class (interval) will we include them? Can we include 35.5 in class 31-35? In class 36-40?

We can not do so. The class 31-35 includes numbers upto 35 and the class 36-40, includes numbers from 36 onwards. So, there are gaps in between the upper and lower limits of two consecutive classes. To overcome this difficulty, we divide the intervals in such a way that the upper and lower limits of consecutive classes are the same. For this, we find the difference between the upper limit of a class and the lower limit of its succeeding class. We than add half of this difference to each of the upper limits and subtract the same from each of the lower limits. For example

Consider the classes 31 – 35 and 36-40

The lower limit of 36 – 40 is 36

The upper limit of 31 – 35 is 35

The difference = 36 – 35 = 1

So, half the difference = $\frac{1}{2} = 0.5$

So, the new class interval formed from 31-35 is $(31 - 0.5) - (35 + 0.5)$, i.e., 30.5 - 35.5. Similarly, class 36-40 will be $(36 - 0.5) - (40 + 0.5)$, i.e., 35.5 – 40.5 and so on.

This way, the new classes will be

30.5-35.5, 35.5-40.5, 40.5-45.5, 45.5-50.5, 50.5-55.5, 55.5-60.5, 60.5-65.5, 65.5-70.5 and 70.5-75.5. These are now continuous classes.

These changed limits are called **true class limits**. Thus, for the class 30.5-35.5, 30.5 is the **true lower class limit** and 35.5 is the **true upper class limit.**

Now obviously, 35.5 will be included in the class 35.5-40.5 and 50.54 in the class 50.5-55.5.

So, the new frequency distribution will be as follows:

Weight (in kg)	Number of Students	
30.5-35.5	10	
35.5-40.5	8 ←	35.5 included in the class
40.5-45.5	15	
45.5-50.5	4	
50.5-55.5	3 ←	50.54 included in the class
55.5-60.5	3	
60.5-65.5	4	
65.5-70.5	3	
70.5-75.5	2	
Total	**52**	

ILLUSTRATIVE EXAMPLES

Example .1 : The heights of 30 students, (in centimetres) have been found to be as follows:

161 151 153 165 167 154
162 163 170 165 157 156
153 160 160 170 161 167
154 151 152 156 157 160
161 160 163 167 168 158

Represent the data by a grouped frequency distribution table, taking the classes as 161-165, 166-170, etc.

Solution:

(i) Frequency distribution table showing heights of 30 students

Height (in cm)	Tally	Marks Frequency
151-155	𝍸 ||	7
156-160	𝍸 ||||	9
161-165	𝍸 |||	8
166-170	𝍸 |	6
Total		30

Example .2: Construct a frequency table for the following data which give the daily wages (in rupees) of 32 persons. Use class intervals of size 10.

110 184 129 141 105 134 136 176 155
145 150 160 160 152 201 159 203 146
177 139 105 140 190 158 203 108 129
118 112 169 140 185

Solution: Range of data = 205 – 105 = 98

Frequency distribution table of the above data is given below:

Frequency table showing the daily wages of 32 persons

Daily wages (in Rs.)	Tally Marks	Number of persons or frequency
105-115	𝍸	5
115-125	|	1
125-135	|||	3
135-145	𝍸	5
145-155	||||	4
155-165	𝍸	5
165-175	|	1
175-185	|||	3
185-195	||	2
195-205	|||	3
Total		32

EXERCISE 8.1

1. Heights (in cm) of 30 boys in Class IX are given below:

140 140 160 139 153 146 151 150 150 154
148 158 151 160 150 149 148 140 148 153
140 139 150 152 149 142 152 140 146 148

Determine the range of the data.

2. Following is the frequency distribution of ages (in years) of 40 workers in a factory:

Age (in years)	Number of workers
25-31	12
31-37	15
37-43	7
43-49	5
49-55	1
Total	40

(i) What is the class size?

(ii) What is the upper class limit of class 37-43?

(iii) What is the lower class limit of class 49-55?

3. 30 girls of Class X appeared for a test. The marks obtained by them are given as follows:

46	31	74	68	42	54	14	93	72	53
59	38	16	88	27	44	63	43	81	64
77	62	53	40	71	60	8	68	50	58

Construct a grouped frequency distribution of the data using the classes 0-9, 10-19 etc. Also, find the number of girls who secured marks more than 49.

4. Construct a frequency table with class intervals of equal sizes using 310-330 as one of the class interval for the following data:

268	230	368	248	242	310	272	342
310	300	300	320	315	304	402	316
406	292	355	248	210	240	330	316
406	215	262	238				

ANSWERS 8.1

1. 21 cm
2. (a) 6 (b) 43 (c) 49

3.

Marks	Number of students
0-10	1
10-19	2
20-29	1
30-39	2
40-49	5
50-59	6
60-69	6
70-79	4
80-89	2
90-99	1
Total	30

4.

Class interval	Frequency
210-230	2
230-250	5
250-270	2
270-290	2
290-310	4
310-330	6
330-350	2
350-370	2
370-390	0
390-410	3
Total	25

19 girls secured more than 49 marks.

8.4 CUMULATIVE FREQUENCY DISTRIBUTION

Consider the following frequency distribution table:

Weight (in kg)	Number of Students
30-35	10
35-40	7
40-45	15
45-50	4
50-55	2
55-60	3
60-65	4
65-70	3
70-75	2
Total	**50**

Now can you answer the following questions:

(i) How many students have their weights less than 35 kg?

(ii) How many students have their weights less than 50 kg?

(iii) How many students have their weights less than 60 kg?

(iv) How many students have their weights less than 70 kg?

Let us try to find the answers :

Number of students with weight:

Less than 35 kg : 10

Less than 40 kg : (10) + 7 = 17

Less than 45 kg : (17) + 15 =32

Less than 50 kg : (32) + 4 = 36

Less than 55 kg : (36) + 2 = 38

Less than 60 kg : (38) + 3 = 41

Less than 65 kg : (41) + 4 = 45

Less than 70 kg : (45) + 3 = 48

Less than 75 kg : (48) + 2 = 50

The frequencies 10, 17, 32, 36, 38, 41, 48, 50 are called the **cumulative frequencies** of the respective classes. The cumulative frequency of the last class, i.e., 70-75 is 50 which is the total number of observations.

In the table we insert a column showing the cumulative frequency of each class, and get cumulative frequency distribution table of the data.

Cumulative Frequency Distribution Table

Weight (in kg)	Number of students (frequency)	Cumulative frequency
0-35	10	10
35-40	7	17
40-45	15	32
45-50	4	36
50-55	2	38
55-60	3	41
60-65	4	45
65-70	3	48
70-75	2	50
Total	50	

EXERCISE 8.2

1. Construct a cumulative frequency distribution for each of the following distributions:

(i)

Classes	Frequency
1-5	4
6-10	6
11-15	10
16-20	13
21-25	6
26-30	2

(ii)

Classes	Frequency
0-10	3
10-20	10
20-30	24
30-40	32
40-50	9
50-60	7

2. Construct a cumulative frequency distribution from the following data:

Heights (in cm)	110-120	120-130	130-140	140-150	150-160	Total
Number of students	14	30	60	42	14	160

How many students have their heights less than 150 cm?

ANSWER 8.2

1. (i)

Classes	Frequency	Cumulative frequency
1-5	4	4
6-10	6	10
11-15	10	20
16-20	13	33
21-25	6	39
26-30	2	41
Total	41	

(ii)

Classes	Frequency	Cumulative frequency
0-10	3	3
10-20	10	13
20-30	24	37
30-40	32	69
40-50	9	78
50-60	7	85
Total	85	

2.

Heights (in cm)	Number of students	Cumulative frequency
110-120	14	14
120-130	30	44
13-140	60	104
140-150	42	146
150-160	14	160
Total	160	

140 students have heights less than 150.

8.5 Graphical Representation of Data

Bar Graphs

We have discussed presentation of data by tables. There is another way to present the data càlled graphical representation which is more convenient for the purpose of comparison among the individual items. For example Fig 10.1 represents the data given in the table regarding blood groups.

Blood groups of 35 students in a class

Blood Group	Number of students
A	13
B	9
AB	6
O	7
Total	35

We can represent this data by Fig. 10.1

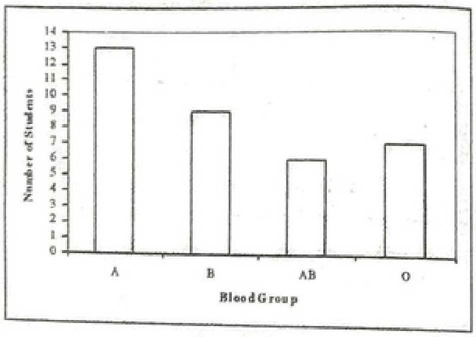


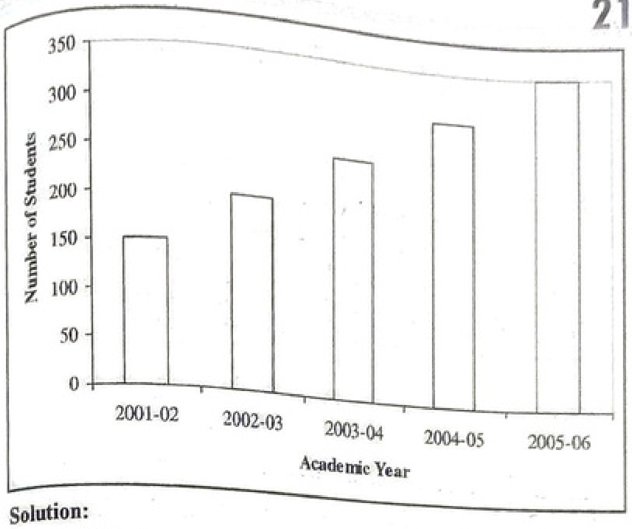
Fig. 10.1

This is called a bar graph.

Bars (rectangles) of unifoirm width are drawn with equal spaces in between them, on the horizontal axis-called x-axis. The heights of the rectangles are shown along the vertical axis-known as y-axis and are proportional to their respective frequencies.

Example .3: Given below is the bar graph of the number of students in Class IX during academic years 2001-02 to 2005-06. Read the bar graph and answer the following questions:

(i) In which year is the number of students in the class, 250?



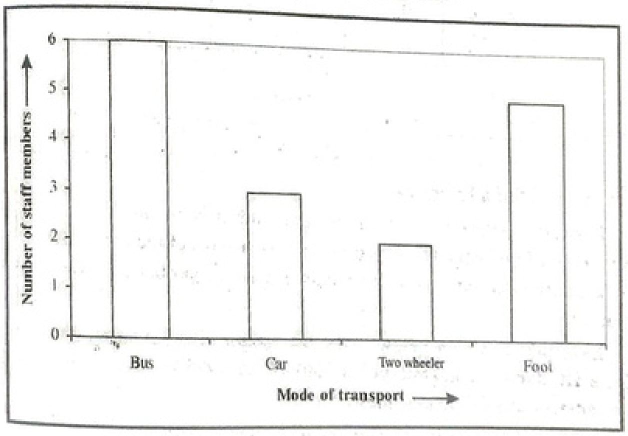
Solution:

(i) In 2003-04, the number of students in the class was 250.

Exercise 8.3

1. The following bar graph shows how the members of the staff of an office come to office.

Mode of transport of office staff



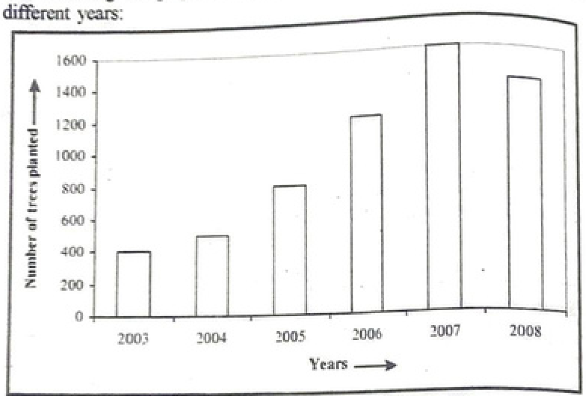
Study the bar graph and answer the following questions:

(i) How many members of staff come to office on two wheeler?

(ii) How many member of staff come to office by bus?

(iii) What is the most common mode of transfport of the members of staff?

2. The following bar graph shows the number of trees planted by an agency in different years:



Study the above bar graph and answer the following questions:

(i) What is the total number of trees planted by the agency from 2003 to 2008?

(ii) In which year is the number of trees planted the maximum?

(iii) In which year is the number of trees planted the minimum?

(iv) In which year, the number of trees planted is less than the number of trees planted in the year preceding it?

3. The expenditure of a company under different heads (in lakh of rupees) for a year is given below:

Head	Expenditure (in lakhs of rupees)
Salary of employees	200
Travelling allowances	100
Electricity and water	50
Rent	125
Others	150

Construct a bar chart to represent this data.

ANSWERS 8.3

1. (i) 2 (ii) 6 (iii) Bus
2. (i) 5900 (ii) 2007 (iii) 2003 (iv) 2008

8.6 HISTOGRAMS AND FREQUENCY POLYGONS

We have learnt to represent a given information using a bar graph. Now, we will learn how to represent a grouped frequency distribution graphically. A continuous grouped frequency distribution can be represented graphically by a **histogram**. A **histogram is a vertical bar graph without any space between the bars.**

(i) The classes of the grouped data are taken along the horizontal axis and

(ii) the respective class frequencies on the vertical axis.

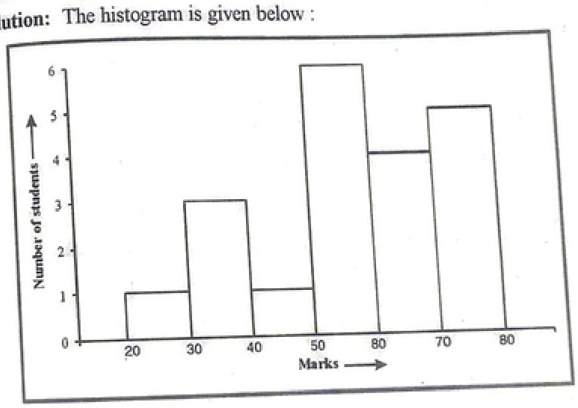
(iii) For each class a rectangle is constructed with base as the width of the class and height determined from the class frequencies.

Example .4: The following is the frequency distribution of marks obtained by 20 students in a class test.

Marks obtained	20-30	30-40	40-50	50-60	60-70	70-80
Number of students	1	3	1	6	4	5

Draw a histogram for the above data.

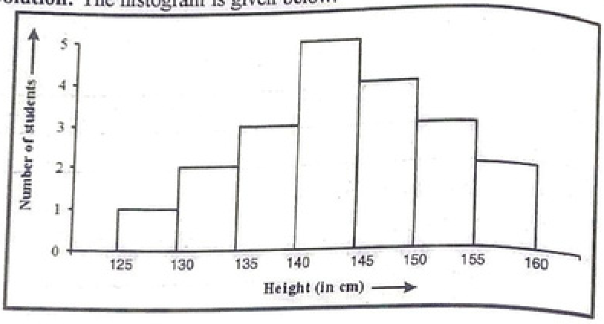
Solution: The histogram is given below :



Example 5: Draw a histogram for the following data:

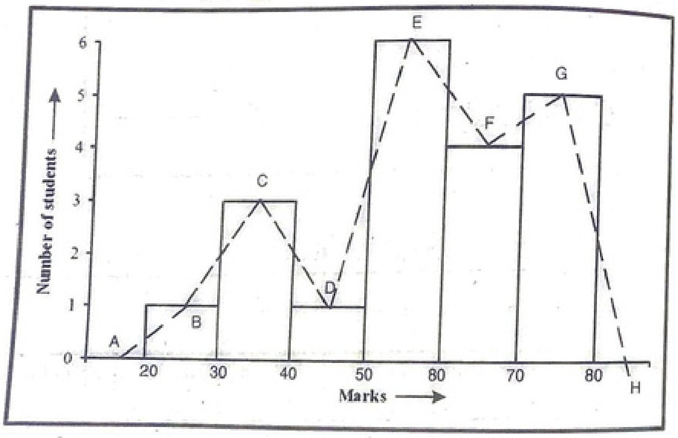
Height (in cm)	125-130	130-135	135-140	140-145	145-150	150-155	155-160
Number of students	1	2	3	5	4	3	2

Solution: The histogram is given below:



8.7 FREQUENCY POLYGON

There is another way of representing a grouped frequency distribution graphically. This is called **frequency polygen**. Consider the following example to under stand the concept of frequency polygon.



Let B, C, D, E, F and G be the mid points of the tops of the adjacent rectangles. Join B to C, C to D, D to E, E to F and F to G by means of line segments (dotted).

To complete the polygon, join B to A (the mid point of class 10-20) and join G to H (the mid point of the class 80-90).

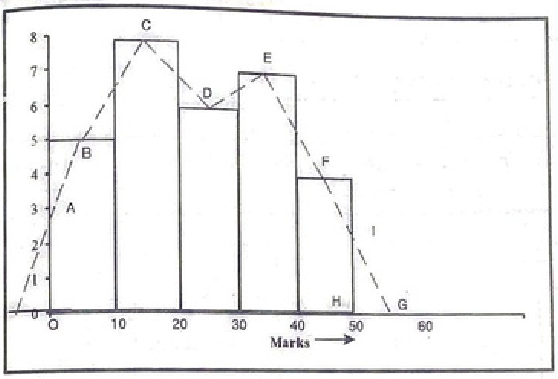
Thus, A B C D E F G H is the **frequency polygon** representing the data given in Example 9

Example 6: Marks (out of 50) obtained by 30 students of Class IX in a Mathematics test are given in the following table:

Marks	0-10	10-20	20-30	30-40	40-50
Number of students	5	8	6	7	4

Draw a frequency polygon for this data.

Solution: First we draw the histogram and then according to above procedure the frequency polygon will be as follows :



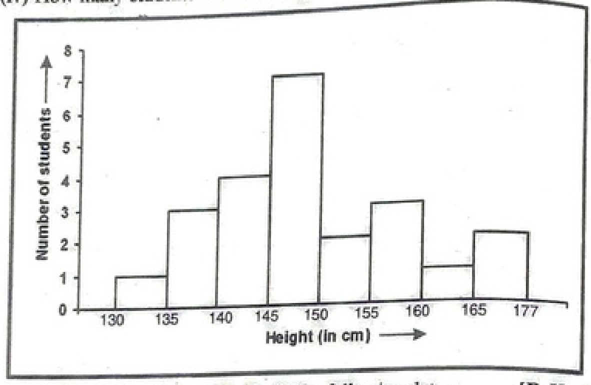
EXERCISE 8.4

1. The daily earnings of 26 workers are given below:

Daily earnings (in ₹)	150-200	200-250	250-300	300-350	350-400
Number of workers	4	8	5	6	3

Draw a histogram to represent the data.

2. Observe the histogram given below and answer the following questions:
   (i) What information is given by the histogram?
   (ii) In which class (group) is the number of students maximum?
   (iii) How many students have the height of 145 cm and above?
   (iv) How many students have the height less than 140 cm?



12. Find median and mode graphically forthe following data : **[R.U. 2016]**

Marks	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100
No. of students	6	4	8	10	4

ANSWER 8.4

2. (i) Heights (in cm) of students (ii) 145 - 150
   (iii) 15 (iv) 4 (v) 4

■■■

Chapter 9 Measures of Central Tendency

9.1 INTRODUCTION

In many problems in statistics we need average of observations; the mid value of the observation or the most repeating observation. These are known as measures of central tendency which we shall study in the present chapter.

9.2 ARITHMETIC AVERAGE OR MEAN

Mean (Arithmetic Average) of Raw Data

To calculate the mean of raw data, all the observations of the data are added and their sum is divided by the number of observations. Thus, the mean of n observations $x_1, x_2, \ldots x_n$ is

$$\frac{x_1 + x_2 + \ldots + x_n}{n}$$

It is generally denoted by $\bar{x}$, so

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

$$= \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(I)}$$

where the symbol "Σ" is the capital letter 'SIGMA' of the Greek alphabet and is used to denote summation. To economise the space required in writing such lengthy expression, we use the symbol Σ, read as **sigma**.

In $\sum_{i=1}^{n} x_i$, i is called the index of summation.

ILLUSTRATIVE EXAMPLES

Example .1: The enrolment in a school in last five years was 605, 710, 745, 835 and 910. What was the average enrolment per year?

Solution: Average enrolment (or mean enrolment)

$$\frac{605+710+745+835+910}{5}=\frac{3805}{5}=761$$

Example .2:The following are the marks in a Mathematics Test of 30 students of Class IX in a school:

40	73	49	83	40	49	27	91	37	31
91	40	31	73	17	49	73	62	40	62
49	50	80	35	40	62	73	49	31	28

Find the mean marks.

Solution: Here, the number of observation (n) = 30

From the Formula (I), the mean marks of students is given by

$$\text{Mean} = (\bar{x}) = \frac{\sum_{i=1}^{30} x_i}{n} = \frac{40+73+...+28}{30}=\frac{1455}{30}=48.5$$

Example 3 . The weight of four bags of wheat (in kg) are 103, 105, 102, 104. Find the mean weight.

Solution: Mean weight ($\bar{x}$) = $\frac{103+105+102+104}{4}kg$

$$=\frac{414}{4}kg = 103.5\ kg$$

Example 4: The mean of 6 observations was found to be 40. Later on, it was detected that one observation 82 was misread as 28. Find the correct mean.

Solution: Mean of 6 observations = 40

So, the sum of all the observations = 6 × 40 = 240

Since one observation 82 was misread as 28,

therefore, correct sum of all the observations = 240 − 28 + 82 = 294

Hence, correct mean = $\frac{294}{6}$ = 49

Example 5: The mean of marks obtained by 30 students of Section A of Class X is 48, that of 35 students of Section B is 50. Find the mean marks obtained by 65 students in Class X.

Solution: Mean marks of 30 students of Section A = 48

So, total marks obtained by 30 students of Section A = 30 × 48 = 1440

Similarly, total marks obtained by 35 students of Section B = 35 × 50 = 1750

Total marks obtained by both sections = 1440 + 1750 = 3190

Mean of marks obtained by 65 students = $\frac{3190}{65}$ = 49.1 approx.

EXERCISES 9.1

1. Find the mean of first ten natural numbers.
2. The heights of 10 girls were measured in cm and the results were as follows:
   142, 149, 135, 150, 128, 140, 149, 152, 138, 145
   Find the mean height.
3. The daily sale of sugar for 6 days in a certain grocery shop is given below. Calculate the mean daily sale of sugar.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
74 kg	121 kg	40 kg	82 kg	70.5 kg	130.5 kg

4. The maximum daily temperature (in oC) of a city on 12 consecutive days are given below:
   32.4 29.5 26.6 25.7 23.5 24.6
   24.2 22.4 24.2 23.0 23.2 28.8
   Calcualte the mean daily temperature.
5. The mean marks obtained by 25 students in a class is 35 and that of 35 students is 25. Find the mean marks obtained by all the students.
6. Mean of 9 observatrions was found to be 35. Later on, it was detected that an observation which was 81, was taken as 18 by mistake. Find the correct mean of the observations.

ANSWERS 9.1

1. 5.5 2. 142.8 cm 3. 86.33 kg
4. 25.8 °C 5. 29.17 6. 42

9.3 Mean of Ungrouped Data

The procedure to find mean of ungrouped data is explained through the following example :

Find the mean of the marks (out of 15) obtained by 20 students.

12 10 5 8 15 5 2 8 10 5
10 12 12 2 5 2 8 10 5 10

Frequency table of the data is :

Marks (x_i)	Number of students (f_i)
2	4
5	5
8	3
10	5
12	2
15	1
	$\Sigma f_i = 20$

Marks (x_i)	Number of students (f_i)	$f_i x_i$
2	4	$2 \times 4 = 8$
5	5	$5 \times 5 = 25$
8	3	$3 \times 8 = 24$
10	5	$5 \times 10 = 50$
12	2	$2 \times 12 = 24$
15	1	$1 \times 15 = 15$
	$\Sigma f_i = 20$	$\Sigma f_i x_i = 146$

Now required formula is $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{146}{20} = 7.3$$

Illustrative Examples

Example 1: The following data represents the weekly wages (in rupees) of the employees:

Weekly wages (in ₹)	900	1000	1100	1200	1300	1400	1500
Number of employees	12	13	14	13	14	11	5

Find the mean weekly wages of the employees.

Weekly wages (in ₹) (x_i)	Number of employees (f_i)	$f_i x_i$
900	12	10800
1000	13	13000
1100	14	15400
1200	13	15600
1300	12	15600
1400	11	15400
1500	5	7500
	$\Sigma f_i = 80$	$\Sigma f_i x_i = 93300$

Using the Formula $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

$$\text{Mean weekly wages} = \frac{\sum f_i x_i}{\sum f_i} = \text{Rs. } \frac{93300}{80} = \text{Rs. } 1166.25$$

Sometimes when the numerical values of x_i and f_i are large, finding the product f_i and x_i becomes difficult and time consuming.

Here we introduce a **short-cut method.** We choose an arbitrary constant a, also called the **assumed mean** and subtract it from each of the values x_i. The reduced value, $d_i = x_i - a$ is called the **deviation** of x_i from 'a'.

and are use following formula

$$\bar{x} = a + \frac{1}{N}\sum f_i d_i$$

where $N = \Sigma f_i$

This method of calcualtion of mean is known as **Assumed Mean Method.**

Let $a = 1200$.

Weekly wages (in ₹) (x_i)	Number of employees (f_i)	Deviations $d_i = x_i - 1200$	$f_i d_i$
900	12	– 300	– 3600
1000	13	– 200	– 2600
1100	14	– 100	– 1400
1200	13	0	0
1300	12	100	+ 1200
1400	11	200	+ 2200
1500	5	300	+ 1500
	$\Sigma f_i = 80$		$\Sigma f_i d_i = -2700$

Using Formula

$$\bar{x} = a + \frac{1}{N}\sum f_i d_i$$

$$= 1200 + \frac{1}{80}(-2700)$$

$$= 1200 - 33.75 = 1166.25$$

So, the mean weekly wages = Rs. 1166.25

Example 2. If the mean of the following data is 20.2, find the value of *k*.

x_i	10	15	20	25	30
f_i	6	8	20	k	6

Solution : Mean $= \dfrac{\sum f_i x_i}{\sum f_i} = \dfrac{60+120+400+25k+180}{40+k}$

$$= \frac{760+25k}{40+k}$$

So, $\dfrac{760+25k}{40+k} = 20.2$ (Given)

or $760 + 25k = 20.2(40 + k)$

or $7600 + 250k = 8080 + 202k$

or $k = 10$

EXERCISES 9.2

1. The wieghts (in kg) of 70 teachers in a factory are given below. Find the mean weight of a teacher.

Weight (in kg)	Number of Teachers
60	10
61	8
62	14
63	16
64	15
65	7

2. If the mean of following data is 17.45 determine the value of k:

x	15	16	17	18	19	20
f	3	8	10	k	5	4

3. Calcualte the mean for each of the following distributions:

(i)

x	6	10	15	18	22	27	30
f	12	36	54	72	62	42	22

(ii)

x	5	5.4	6.2	7.2	7.6	8.4	9.4
f	3	14	28	23	8	3	1

4. Find the mean marks of the following distribution:

Marks	1	2	3	4	5	6	7	8	9	10
Frequency	1	3	5	9	14	18	16	9	3	2

ANSWERS 9.2

1. 11.68 2. 10
3. (i) 18.99 (ii) 6.57 4. 5.84

9.4 MEAN OF GROUPED DATA

Consider the following grouped frequency distribution:

Daily wages (in ₹)	Number of workers
150-160	5
160-170	8
170-180	15
180-190	10
190-200	2

To find mean of the grouped frequency distribution, we find mid value (class mark) of each interal and proceed as follows :

Daily wages (in ₹)	Number of workers (f_i)	Class marks (x_i)	$f_i x_i$
150-160	5	155	775
160-170	8	165	1320
170-180	15	175	2625
180-190	10	185	850
190-200	2	195	390
	$\Sigma f_i = 40$		$\Sigma f_i x_i = 6960$

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{6960}{40} = 174$$

So, the mean daily wage = ₹ 174

This method of calculating the mean of grouped data is Direct Method. We can also find the mean of grouped data by using Assumed Mean Method as follows:

Let assumed mean $a = 175$

Daily wages (in ₹)	Number of workers (f_i)	Class marks (x_i)	Deviations $d_i = x_i - 175$	$f_i d_i$
150-160	5	155	−20	−100
160-170	8	165	−10	−80
170-180	15	175	0	0
180-190	10	185	+10	100
190-200	2	195	+20	40
	$\Sigma f_i = 40$			$\Sigma f_i d_i = -40$

So, using formula,

$$\bar{x} = a + \frac{1}{N} \sum f_i d_i$$

$$= 175 + \frac{1}{40}(-40)$$

$$= 175 - 1 = 174$$

Thus, the mean daily wage = ₹ 174

ILLUSTRATIVE EXAMPLE

Example 1: Find the mean for the following frequency distribution by (i) Direct Method, (ii) Assumed Mean Method. (ii) Step deviation method

Class	Frequency
20-40	9
40-60	11
60-80	14
80-100	6
100-120	8
120-140	15
140-160	12
Total	75

Solution : (i) Direct Method

Class	Frequency (f_i)	Class marks (x_i)	$f_i x_i$
20-40	9	30	270
40-60	11	50	550
60-80	14	70	980
80-100	6	90	540
100-120	8	110	880
120-140	15	130	1950
140-160	12	150	1800
	$\Sigma f_i = 75$		$\Sigma f_i x_i = 6970$

$$\text{So, mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{6960}{75} = 92.93$$

(ii) Assumed mean method :

Let assumed mean = a = 90

Class	Frequency (f_i)	Class marks (x_i)	Deviation $d_i = x_i - 90$	$f_i d_i$
20-40	9	30	− 60	− 540
40-60	11	50	− 40	− 440
60-80	14	70	− 20	− 280
80-100	6	90	0	0
100-120	8	110	+ 20	160
120-140	15	130	+ 40	600
140-160	12	150	+ 60	720
	N = Σf_i = 75			$\Sigma f_i d_i$ = 220

$$\bar{x} = a + \frac{1}{N}\sum f_i d_i = 90 + \frac{220}{75} = 92.93$$

(iii) Step - Deviation method

In the table above, the class marks are all multiples of 20. So, if we divide these value by 20, and get smaller numbers to multiply with f_i.

Let $u_i = \frac{x_i - a}{h}$, where a is the assumed mean and h is the class size.

Now we can find mean by using the formula

$$\text{Mean} = \bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \qquad \text{(IV)}$$

Take $a = 90$, Here $h = 20$

Class	Frequency (f_i)	Class marks (x_i)	Deviation $d_i = x_i - 90$	$u_i = \frac{x_i - a}{h}$	$f_i u_i$
20-40	9	30	− 60	− 3	− 27
40-60	11	50	− 40	−2	− 22
60-80	14	70	− 20	− 1	− 14
80-100	6	90	0	0	0
100-120	8	110	+ 20	1	8
120-140	15	130	+ 40	2	30
140-160	12	150	+ 60	3	36
	Σf_i = 75				$\Sigma f_i u_i$ = 11

Using the formula

$$\bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \times h = 90 + \frac{11}{75} \times 20$$

$$= 90 + \frac{220}{75} = 92.93$$

(Note : Calculating mean by using above Formula is known as Step-deviation **Method.)**

Example 2: Calcualte the mean daily wage from the following distribution by using Step deviation method.

Daily wages (in ₹)	150-160	160-70	170-180	180-190	190-200
Numbr of workers	5	8	15	10	2

Solution:

Let $a = 175$. Here $h = 10$

Daily wages (in ₹)	Number of workers (f_i)	Class marks (x_i)	Deviation $d_i = x_i - 90$	$u_i = \frac{x_i - a}{h}$	$f_i u_i$
150-160	5	155	− 20	− 2	− 10
160-170	8	165	− 10	− 1	− 8
170-180	15	175	0	0	0
180-190	10	185	10	1	10
190-200	2	195	20	2	4
	Σf_i = 40				$\Sigma f_i u_i$ = − 4

Using formula

$$\bar{x} = a + \left(\frac{\sum f_i U_i}{\sum f_i}\right) \times h = 175 + \frac{-4}{40} \times 10 = ₹174$$

EXERCISES 9.3

1. The following is the distribution of bulbs kept in boxes :

Number of bulbs	50-52	52-54	54-56	56-58	58-60
Number of boxes	15	100	126	105	30

Find the mean number of bulbs kept in a box.

2. Following table shows marks obtained by 100 students in a mathematics test

Marks	0-10	10-20	20-30	30-40	40-50	50-60
Number of students	12	15	25	25	17	6

3. Find the mean of the following data by using (i) Assumed Mean Method and (ii) Step deviation Method.

Class	150-200	200-250	250-300	300-350	350-400
Frequency	48	32	35	20	10

4. The weekly observations on cost of living index in a certain city for a particular year are given below:

Cost of living index	140-150	150-160	160-170	170-180	180-190	190-200
Number of weeks	5	8	20	9	6	4

Calculate mean weekly cost of living index by using Step deviation Method.

ANSWER 9.3

1. 55.19 2. 28.80 3. 244.66 4. 167.9

9.5 MEDIAN

Median is a measure of central tendency which gives the value of the middle most observation in the data when the data is arranged in ascending (or descending) order.

9.5.1 MEDIAN OF RAW DATA

Median of raw data is calculated as follows :

(i) Arrange the (numerical) data in an ascending (or descending) order.

(ii) When the number of observations (n) is **odd**, the median is the value of $\left(\frac{n+1}{2}\right)$th observation.

(iii) When the number of observation (n) is even, the median is the mean of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th observations.

The methods to find median of ungrouped data and grouped data are explained in the following examples.

Example 3. The weights (in kg) of 15 dogs are as follows :

9, 26, 10, 22, 36, 13, 20, 20, 10, 21, 25, 16, 12, 14, 19

Find the median weight.

Solution : Let us arrange the data in the **ascending (or descending)** order :

9, 10, 10, 12, 13, 14, 16, 19, 20, 20, 21, 22, 25, 36
↑

Here, number of observations = 15

So, the median will be $\left(\frac{n+1}{2}\right)$th, i.e., $\left(\frac{15+1}{2}\right)$th, i.e. 8th observation which is 19 kg.

Example 4. The points scored by a basket ball team in a series of matches are as follows :

16, 1, 6, 16, 14, 4, 13, 8, 9, 23, 47, 9, 7, 8, 17, 28

Find the median of the data.

Solution : Here number of observations = 16

So, the median will be the mean of $\left(\frac{16}{2}\right)$th and $\left(\frac{16}{2}+1\right)$th, i.e. mean of 6th and 9th observations, when the data is arranged in ascending (or desceding) order as :

1, 4, 6, 7, 8, 8, 9, 9, 13, 14, 16, 17, 23, 26, 28, 47
↑ 8th term ↑ 9th term

So, the median $= \frac{9+13}{2} = 11$

9.5.2 MEDIAN OF UNGROUPED DATA

We illustrate calculation of the median of ungrouped data through examples.

Example 5. Find the median of the following data, which gives the marks, out of 15, obtaine by 35 students in a mathematics test.

Marks obtained	3	5	6	11	15	14	13	7	12	10
Number of Students	4	6	5	7	1	3	2	3	3	1

Solution : First arrange marks in ascending order and prepare a frequency table as follows :

Marks obtained	3	5	6	7	10	11	12	13	14	15
Number of Students (frequency)	4	6	5	3	1	7	3	2	3	1

Here $n = 35$, which is odd. So, the median will be $\left(\frac{n+1}{2}\right)$th i.e. $\left(\frac{35+1}{2}\right)$th, i.e. 18th observation.

To find value of 18th observation, we prepare cumulative frequency table as follows :

Marks obtained	Number of students	Cumulative frequency
3	4	4
5	6	10
6	5	15
7	3	18
10	1	19
11	7	26
12	3	29
13	2	31
14	3	34
15	1	35

From the table above, we see that 18th observation is 7

So, Median = 7

■ **Example 6 : Find the median of the following data :**

Weight (in kg)	40	41	42	43	44	45	46	48
Number of students	2	5	7	8	13	26	6	3

Solution : Here $n = 2 + 5 + 7 + 8 + 13 + 26 + 6 + 3 = 70$, Which is even, and weight are already arranged in the ascending order. Let us prepare cumulative frequency table of the data :

Weight (in kg)	Number of students (frequency)	Cumulative frequency	
40	2	2	
41	5	7	
42	7	14	
43	8	22	
44	13	35	35th observation
45	26	61	36th observation
46	6	67	
48	3	70	

Since n is even, so the median will be the mean of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th observations, i.e., 35th and 36th observation. From the table, we see that

35 the observation is 44

and 36th observation is 45

So, Median $= \frac{44+45}{2} = 44.5$

9.5.3 MEDIAN OF GROUPED DATA

In a grouped data, we may not be able to find the middle observation by looking at the cumulative frequencies as the middle observation will be some value in a class interval. It is, therefore, necessary to find the value inside a class that divides the whole distribution into two halves.

To find this class, we find the cumulative frequencies of all the classes and $\frac{N}{2}$. We now locate the class whose cumulative frequency is greater than (and nearest to) $\frac{N}{2}$. This is called the median class.

After finding the median class, we use the following formula for calculating the median.

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

where l = lower limit of median class,

N = number of observations,

cf = cumulative frequency of class preceding the median class,

f = frequency of median class,

h = class size (assuming class size to be equal).

■ **Example 7 : The distribution below gives the weights of 30 students of a class. Find the median weight of the students.**

Weight (in Kg)	40-45	45-50	50-55	55-60	60-65	65-70	70-75
Number of Students	2	3	8	6	6	3	2

Solution :

Weight (in Kg)	40-45	45-50	50-55	55-60	60-65	65-70	70-75
Number of Students	2	3	8	6	6	3	2
Cumulative Frequency c.f.	2	5	13	19	25	28	30

Clearly $\frac{N}{2} = \frac{30}{2} = 15$

C.F. just greater than 15 is 19, therfore median class is 55 to 60.

Here $l = 55$, $\frac{N}{2} = \frac{30}{2} = 15$, c = 13, f = 6, h = 5

$$\therefore \quad \text{Median} = l + \left(\frac{\frac{N}{2} - cf}{f}\right) \times h$$

$$= 55 + \left(\frac{15-13}{6}\right) \times 5 = 55 + \frac{2}{6} \times 5 = 55 + \frac{5}{3} = 55 + 1.67$$

$\therefore$ Medial $= 56.67$

Hence, the median weight of the students is 56.67 kg.

EXERCISES-9.4

1. Following are the goals scored by a team in a series of 11 matches
1, 0, 3, 2, 4, 5, 2, 4, 4, 2, 5
Determine the median score.

2. In a diagnostic test in mathematies given to 12 students, the following marks (out of 100) are recorded
46, 52, 48, 39, 41, 62, 55, 53, 96, 39, 45, 99
Calculate the median for this data.

3. A fair die is thrown 100 times and it outcomes are recorded as shown below :

Outcome	1	2	3	4	5	6
Frequency	17	15	16	18	16	18

Find the median outcome of the distributions.

4. For each of the following frequency distribution, find the median :

(a)

x_i	2	3	4	5	6	7
f_i	4	9	16	14	11	6

(b)

x_i	5	10	15	20	25	30	35	40
f_i	3	7	12	20	28	31	28	26

(c)

x_i	2.3	3	5.1	5.8	7.4	6.7	4.3
f_i	5	8	14	21	13	5	7

5. The following frequency distribution gives the monthly consumption of electricity of 68 consumers of a locality. Find the median and mean of the data and compare them.

Monthly consumption (in units)	Number of consumers
65 - 85	4
85 - 105	5
105 - 125	13
125 - 145	20
145 - 165	14
165 - 185	8
185 -205	4

6. If the median of the distribution given below is 28.5, find the values of x and y.

Class interval	Frequency
0 - 10	5
10 - 20	x
20 - 30	20
30 - 40	15
40 - 50	y
50 - 60	5
Total	60

7. The following table gives the distribution of the life time of 400 neon lamps :

Life time (in hours)	Number of lamps
1500 - 2000	14
2000 - 2500	56
2500 - 3000	60
3000 - 3500	86
3500 - 4000	74
4000 - 4500	62
4500 - 5000	48

Find the median life time of a lamp.

ANSWERS 9.4

1. 3 2. 50 3. 4

4. (a) 4 (b) 30 (c) 5.8
5. 137, 137.05
6. $x = 8, y = 7$
7. 3406.98 hours

9.6 MODE

Look at the following example:

A company produces readymade shirts of different sizes. The company kept record of its sale for one week which is given below:

size (in cm)	90	95	100	105	110	115
Number of shirts	50	125	190	385	270	28

From the table, we see that the sales of shirts of size 105 cm is maximum. So, the company will go ahead producing this size in the largest number. Here, 105 is nothing but the mode of the data. Mode is also one of the measures of central tendency.

The observation that occurs most frequently in the data is called mode of the data.

In other words, the observation with maximum frequency is called mode of the data.

The readymade garments and shoe industries etc, make use of this measure of central tendency. Based on mode of the demand data, these industries decide which size of the product should be produced in large numbers to meet the market demand.

9.6.1 MODE OF RAW DATA

In case of raw data, it is easy to pick up mode by just looking at the data. Let us consider the following example:

■ **Example 8 : The number of goals scored by a football team in 12 matches are :**

1, 2, 2, 3, 1, 2, 2, 4, 5, 3, 3, 4

What is the modal score ?

Solution : Just by looking at the data, we find the frequency of 2 is 4 and is more than the frequency of all other scores.

So, mode of the data is 2, or modal score is 2.

■ **Example 9 : Find the mode of the data :**

9, 6, 8, 8, 10, 7, 12, 15, 22, 15

Solution : Arranging the data in increasing order, we have

6, 7, 8, 9, 9, 10, 12, 15, 15, 22

We find that the both the observation 9 and 15 have the same maximum frequency 2. So, both are the modes of the data.

Remarks : 1. In this lesson, we will take up the data having a single mode only.

2. In the data, if each observation has the same frequency, then we say that the data does not have a mode.

9.6.2 Mode of Ungrouped Data

Let us illustrate finding of the mode of ungrouped data through an example

■ **Example 10 : Find the mode of the following data :**

Weight (in kg)	40	41	42	43	44	45	46	48
Number of Students	2	6	8	9	10	22	13	5

Solution : From the table, we see that the weight 45 kg has maximum frequency 22 which means that maximum number of students have their weight 45 kg. So, the mode is 45 kg or the modal weight is 45 kg.

9.6.3 Mode of Grouped Data

In a grouped frequency distribution, it is not possible to determine the mode by looking at the frequencies. Here, we can only locate a class with the maximum frequency, called the **modal class.** The mode is a value inside the modal class, and is given by the formula :

$$\text{mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

where l = lower limit of the modal class,

h = size of the calss interval (assuming all calss sized to be qual),

f_1 = frequency of the modal class,

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class.

■ **Example 11 : A survey conducted on 20 households in a locality by a group of students resulted in the following frequency table for the number of family members in a household :**

Family size	1–3	3–5	5–7	7–9	9–11
Number of families	7	8	2	2	1

Find the mode of this data.

Solution : Here the maximum class frequency is 8, and the class corresponding to this frequency is 3-5. So, the modal class is 3-5.

Now

modal class = 3-5, lower limit (l) of modal class = 3, class size (h) = 2

frequency (f_1) of the modal class = 8,

frequency (f_0) of class preceding the modal class = 7,

frequency (f_2) of class succeeding the modal class = 2.

Now, let us substitute these values in the formula :

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$$= 3 + \left(\frac{8-7}{2\times 8-7-2}\right) \times 2 = 3 + \frac{2}{7} = 3.286$$

Therefore, the mode of the data above is 3.286.

Exrcise-9.5

1. The number of TV sets in each of 15 households are found as given below:

   2, 2, 4, 2, 1, 1, 1, 2, 1, 1, 3, 3, 1, 3, 0

   What is the mode of this data?

2. Find the mode of the data:

   5, 10, 3, 7, 2, 9, 6, 2, 11, 2

3. Following are the marks (out of 10) obtained by 80 students in a mathematics test:

Marks obtained	0	1	2	3	4	5	6	7	8	9	10
Number of students	5	2	3	5	9	11	15	16	9	3	2

   Determine the modal marks.

4. A die is thrown 100 times, giving the following results

Outcome	1	2	3	4	5	6
Frequency	15	16	16	15	17	20

Find the modal outcome from this distribution.

5. For the following grouped frequency distribution, find the mode.

Class	3-6	6-9	9-12	12-15	15-18	18-21	21-24
Frequency	2	5	10	23	21	12	3

6. The following table shows the ages of the patients admitted in a hospital during a year.

Age (in years)	5-15	15-25	25-35	35-45	45-55	55-65
No. of patients	6	11	21	23	14	5

Find the mode of the data given above.

ANSWER 9.5

1. 1 2. 2 3. 7 4. 6
5. 14.6 6. 36.82

9.7 STANDARD DEVIATION

Mean deviation is an average of total deviations by ignoring positive and negative signs. We add all the deviations without considering their signs. To correct this mathematical error or contradiction, we use other process to find deviation. Under this process, we calculate arithmetic means and calculate deviations of all variables from this and square all the deviations. Lastly we add all squared numbers and take their average, and then take its square root. The number thus obtained is called standard deviation.

DEFINITIONS

Standard deviation : The square root of the arithmetic mean of the squares of the deviations of the different variate values of series from their arithmetic mean is called standard deviation.

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

Coefficient of standard deviation : Standard deviation is an absolute value. For comparative study of two series, relative measure of standard deviation is used, which is called coefficient of standard deviation. Its formula is as follows :

$$\text{Coefficient of standard deviation} = \frac{\sigma}{\bar{x}} \text{ or } \frac{\text{Mean deviation}}{\text{Mean}}$$

Variance : The mean of squares of deviations from mean is called variance, i.e.

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

Coefficient of variation : The coefficient of standard deviation is calculate for comparison of dispersion of two or more series. Its value is always less than one, i.e., its value comes in decimal or fraction form which is not convenient in prediction. Therefore coefficient of variation is used. The percentage obtained on multiplying the coefficient of standard deviation by 100 is called coefficient of variation. In fact coefficient of variation is the percentage form of coefficient of standard deviation. It may be determind by the following formula :

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100$$

Now we calculate standard deviation for different type of data.

(I) For ungrouped data : If n terms in data are $x_1, x_2, x_3, \ldots, x_n$ respectively and their A.M. is $\bar{x}$, then

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

The easier form of the above formula can be written as follows :

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

■ **Example 12. Five students obtained 23, 46, 16 25 and 20 marks in Mathematics respectively. Find the standard deviation of their obtained marks.**

Solution : A.M. of variate values $\bar{x} = \frac{23+46+16+25+20}{5} = \frac{130}{5} = 26$

x	$x-\bar{x}$	$(x-\bar{x})^2$
23	–3	9
46	20	400
16	–10	100
25	–1	1
20	–6	36
		$\sum(x_i-\bar{x})^2 = 546$

$$\text{Standard deviation } \sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}} = \sqrt{\frac{546}{5}}$$

$$= \sqrt{109.2} = 10.45$$

(II) For ungrouped frequency distribution :

If $x_1, x_2, x_3, \ldots, x_n$ are different values of variables and $f_1, f_2, f_3, \ldots, f_n$ be their frequencies respectively and $\bar{x}$ is its A.M. then

$$\text{Standard deviation } \sigma = \sqrt{\frac{\sum f_i(x_i-\bar{x})^2}{N}}$$

$$\text{Where } \bar{x} = \frac{\sum f_i x_i}{N}, N = \Sigma f_i$$

Example 2. Calculate the standard deviation from the following data :

x	1	12	14	16	18	20	22	24
y	5	8	21	24	18	15	7	2

Solution :

x	f	fx	$(x-\bar{x})$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
10	5	50	–6.5	42.25	211.25
12	8	96	–4.5	20.25	162.00
14	21	294	–2.5	6.25	131.25
16	24	384	–0.5	0.25	6.00
18	18	324	1.5	2.25	40.50
20	15	300	3.5	12.25	183.75
22	8	154	5.5	30.25	211.75
24	2	48	7.5	56.25	112.50
	N = 100	$\Sigma f x_i = 1650$			$\sum f_i(x_i-\bar{x})^2 = 1059.00$

$$\bar{x} = \frac{\sum f_i x_i}{N} = \frac{1650}{100} = 16.50$$

$$\sigma = \sqrt{\frac{\sum f_i(x_i-\bar{x})^2}{N}} = \sqrt{\frac{1059}{100}} = 3.25$$

Short-cut methods for calculating standard deviation :

(i) According to definition

$$\sigma_x^2 = \frac{1}{N}\sum f_i(x_i-\bar{x})^2$$

$$= \frac{1}{N}\sum f_i(x_i^2+\bar{x}^2-2x_i\bar{x})$$

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

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

$$\left[\because N=\sum f_i, \bar{x}=\frac{\Sigma f_i x_i}{N}\right]$$

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

Hence
$$\sigma_x = \sqrt{\frac{1}{N}\sum f_i x_i^2-\left(\frac{1}{N}\sum f_i x_i\right)^2}$$

(ii) If assumed mean = a and let $x_i - a = d_i$, then

$$\sigma_x^2 = \frac{1}{N}\sum f_i(x_i-\bar{x})^2$$

$$= \frac{1}{N}\Sigma f_i(x_i-a+a-\bar{x})^2$$

$$= \frac{1}{N}\Sigma f_i(d_i-d)^2, \text{ where } d_i = x_i - a$$

$$= \frac{1}{N}\Sigma f_i d_i^2-\left(\frac{1}{N}\Sigma f_i d_i\right)^2 = \sigma_d^2$$

$$\sigma_x = \sqrt{\frac{1}{N}\Sigma f_i d_i^2-\left(\frac{1}{N}\Sigma f_i d_i\right)^2} = \sigma_d.$$

(iii) Step deviation method : If class interval is equal in grouped frequency distribution then a common factor equal to class interval is taken out to find deviation from assumed mean. This makes the calculation work easy. Remaining process of calculation is same as earlier.

$$\sigma_x = h \times \sqrt{\frac{1}{N}\Sigma f_i u_i^2 - \left(\frac{1}{N}\Sigma f_i u_i\right)^2}$$

Where $u_i = \frac{x_i - a}{h} = \frac{d_i}{h}$ and $d_i = hu_i$

Mean in this method $\bar{x} = a + h\,\frac{\Sigma f_i u_i}{N}$

Example 3. Calculate the standard deviation, coefficient of standard deviation and coefficient of variation from the following data :

Class	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10
Frequency	2	5	15	7	1

Solution :

Class	Mid values x	Frequency f	fx	fx^2
0 - 2	1	2	2	2
2 - 4	3	5	15	45
4 - 6	5	15	75	375
6 - 8	7	7	49	343
8 - 10	9	1	9	81
		$N = \Sigma f_i = 30$	$\Sigma f_i x_i = 150$	$\Sigma f_i x_i^2 = 846$

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\sum f_i x_i^2}{N} - \left(\frac{\sum f_i x_i}{N}\right)^2}$$

$$= \sqrt{\frac{846}{30} - \left(\frac{150}{30}\right)^2} = \sqrt{28.2 - 25} = 1.79$$

Coefficient standard deviation

$$\text{(C.S.D.)} = \frac{\sigma}{\bar{x}} = \frac{1.79}{5} = 0.36$$

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100 = 0.36 \times 100 = 36$$

Example 4. Find the S.D, C.S.D. and C.V. for the following distribution :

x	9	12	15	18	21	24	27	30
y	20	60	150	250	200	120	50	40

Solution : Let assumed mean $a = 18$

x	f	$d = x - 18$	d^2	fd	fd^2
9	20	–9	81	–180	1620
12	60	–6	36	–360	2160
15	150	–3	9	–450	1350
18	250	0	0	0	0
21	200	3	9	600	1800
24	120	6	36	720	4320
27	50	9	81	450	4050
30	40	12	144	480	5760
	$N = 890$			$\Sigma f d_i = 1260$	$\Sigma f d_i^2 = 21060$

$$\text{Standard deviation } \sigma = \sqrt{\frac{1}{N}\sum f_i d_i^2 - \left(\frac{1}{N}\sum f_i d_i\right)^2}$$

$$= \sqrt{\frac{21060}{890} - \left(\frac{1260}{890}\right)^2}$$

$$= \sqrt{23.66 - 2.004}$$

$$= \sqrt{21.656}$$

$$= 4.65$$

$$\text{Mean } \bar{x} = a + \frac{1}{N}\sum f_i d_i$$

$$= 18 + \frac{1260}{890}$$

$$= 19.41$$

$$\text{Coefficient of standard deviation} = \frac{\sigma}{\bar{x}} = \frac{4.65}{19.14}$$

$$= 0.239$$

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100 = 0.239 \times 100$$

$$= 23.9$$

(III) For grouped frequency distribution : For the grouped frequency distribution of equal class intervals, step deviation method or short-cut method is used to find standard deviation.

Example 5. Find the mean and standard deviation of the following distribution :

Class	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students	5	8	15	16	6

Solution : We obtain the solution of this question from step deviation method. Let assumed mean $a = 25$, which is mid value of class 20-30.

Class	Mid values x	No. of students f	$u = \frac{x-25}{10}$	u^2	fu	fu^2
0-10	5	5	−2	4	−10	20
10-20	15	8	−1	1	−8	8
20-30	25	15	0	0	0	0
30-40	35	16	1	1	16	16
40-50	45	6	2	4	12	24
		$N = 50$		10	$\Sigma f_i u_i = 10$	$\Sigma f_i u_i^2 = 68$

$$\text{Mean } \bar{x} = a + h \times \frac{\sum f_i u_i}{N}$$

$$= 25 + \frac{10 \times 10}{50} = 27$$

$$\text{Standard deviation } \sigma = h \times \sqrt{\frac{1}{N}\sum f_i u_i^2 - \left(\frac{1}{N}\sum f_i u_i\right)^2}$$

$$= 10 \times \sqrt{\frac{68}{50} - \left(\frac{10}{50}\right)^2}$$

$$= 10 \times \sqrt{1.32}$$

$$= 10 \times 1.1489$$

$$= 11.489$$

Example 5. Write the merit and demerit of Mean, Median and Mode ? **[R.U. 2015]**

Sol. : Merits of Arithemetic Mean

- It is easy to calculate and simple to understand.

- It is based on all observations and it can be regarded as representative of the given data.
- It is least affected by the fluctuation of sampling.

Demerits of Arithmetic Mean

- It can either be determined by inspection or by graphical location.
- Arithmetic mean cannot be computed when class intervals have open ends.

Merits of Median

- It is very simple measure of the central tendency of the series. In the case of simple statistical series, just a glance at the data is enough to locate the median value.
- Unlike arithmetic means, median value is not destroyed by the extreme values of the series.

Demerits of Median

Following are the various demerits of median :

- Median fails to be a representative measure in case of such series the different values of which are wide apart from each other, Also, median is of limited representative character as it is not based on all the items in the series.
- When the median is located somewhere between the two middle values, it remains only an approximate measure, not a precise value.

Merits of Mode ;

- Mode is very simple measure of central tendency. Sometimes, just at the series is enough to locate the model value.
- Mode can be located graphically, with the help of histogram.
- Mode is that value which occurs most frequently in the series. Accordingly mode is the best representative value of the series.

Demerits of Mode :

- Mode is an uncertain and vague measure of the central tendency.
- With frequencies of all items identical, it is difficult to identify the modal value.

Example 6. Compute the coefficient of range for the following data : 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21. **[R.U. 2015]**

Solution : Coeff. of range $= \frac{H-L}{H+L}$

$$= \frac{21-11}{21+11} = \frac{10}{32}$$

$$= \frac{5}{16}$$

Example 7. Find the standard deviation of the following data ?

x	12	13	14	15	16	17	18	20
y	4	11	32	21	15	8	6	4

[R.U. 2015]

Solution :

x	y	d = x − 16	d^2	f.d.	f.d^2
12	4	−4	16	−16	64
13	11	−3	9	−33	99
14	32	−2	4	−64	138
15	21	−1	1	−21	21
16	15	0	0	0	0
17	8	1	1	8	8
18	6	2	4	12	24
20	4	4	16	16	64

Standard deviations $\sigma = \sqrt{\frac{1}{N}\Sigma f_i d_i^2 - \left(\frac{\Sigma f_i d_i}{N}\right)^2}$

$= \sqrt{\frac{418}{101} - \left(\frac{-98}{101}\right)^2}$

$= \sqrt{4.139 - 0.9415}$

$= \sqrt{3.19752}$

$= 1.7882$

Example 8. The runs recored in 5 innings by 2 players are as :

A	5	8	10	2	15
B	4	7	9	20	10

Find which player is more consistent **[R.U. 2016]**

Solution : For A, $\bar{x} = \frac{5+8+10+2+15}{5} = 8$

$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{98}{5}} = 7\sqrt{\frac{2}{5}} = 4.42$

∴ Coeff.of standard deviation $= \frac{\sigma}{\bar{x}} = \frac{4.42}{8} = 0.55$

For B; - $\bar{y} = \frac{4+7+9+20+10}{5} = \frac{50}{5} = 10$

$\sigma = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \sqrt{\frac{146}{5}} = 5.403$

Coeff. of standard daviation $= \frac{5.403}{10} = 0.54$

∴ Player B is more consistent.

Example 9. Find coefficient of mean deviation and variance for the following series.

4, 5, 2, 3, 6, 4, 2, 5, 3, 6 **[R.U. 2016]**

Solution :

x	f	fx	$(x-\bar{x})^2$	$f(x-\bar{x})^2$	$f\lvert x-\bar{x}\rvert$
2	2	4	64	128	16
3	2	6	49	98	14
4	2	8	36	72	12
5	2	10	25	50	10
6	2	12	16	32	8
	Σf = 10	Σfx = 40		$\Sigma f(x-\bar{x})^2 = 380$	60

$\bar{x} = \frac{40}{10} = 10$

Variance $= \frac{\Sigma f(x-\bar{x})^2}{\Sigma f} = \frac{380}{10} = 38$

Mean deviation $= \frac{\Sigma f\lvert x-\bar{x}\rvert}{\Sigma f} = \frac{60}{10} = 6$

EXERCISES 9.6

1. Find the standard deviation of the following :

x	5	15	25	35	45	55	65	75
y	3	7	9	23	15	8	6	4

2. Find standard deviation with the help of a standard mean of the following frequency distribution :

Class	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 4
Frequency	2	5	7	13	21	16	8	3

3. Find the variance and standard deviation of the following distribution :

Marks obtained	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students	5	8	15	16	6

4. In the following distribution, find standard deviation and its coefficient.

x	10	12	14	16	18	20	22	24
y	5	8	21	24	18	15	7	2

ANSWERS 9.6

1. 16.819
2. 7.995
3. 132, 11.489
4. 3.25, 0.197

Chapter 10 Correlation and Regression

10.1 Introduction

Correlation analysis is a statistical process in which we determine amount of relation between two or more variables. Analysis of correlation is used also in regression principle. Specially study of these are used in problems of social science, education research policy building and taking important decision etc.

In statistics, principle of correlation is very important. In practical correlation principle is used in salary and life living index and sale and profit etc.

10.2 Definition

Correlation : If the change in one variable results in a direct or inverse (i.e. in opposite direction) change in the other variable, then the relation between them is called correlation.

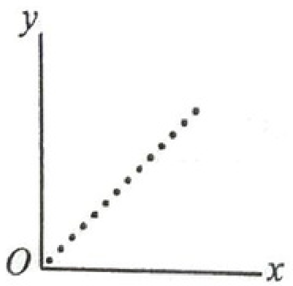
Types of correlation : The correlation based on deviation ratio and number of variable can be classified into three groups. **[R.U. 2016]**

1. On basis of deviation.

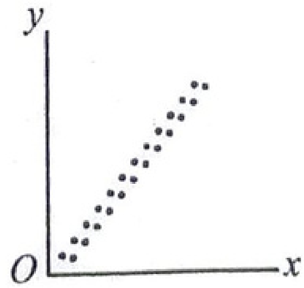
(A) Positive correlation : When corresponding to an increase (or decrease) in one variable there is an increase (or decrease) in the other variable then it is positive correlation between those variates. For example :

(i) Age of husband and age of wife.

(ii) Height of a child and their weight.



Perfect positive correlation

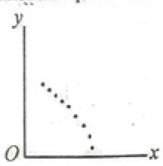


Positive correlation of high degree

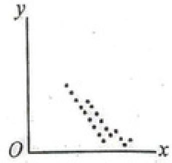
(B) Negative correlation : When corresponding to an increase (or decrease) in one variable there is a decrease (or increase) in other variable i.e., changes in both are in opposite direction, then it is called negative correlation. For example :

(i) Price and demand of an item

(ii) Production and price of an item.



Perfect negative correlation



Negative correlation of high degree

2. On basis of ratio

(A) Linear correlation : If the ratio of changes between two variables be always same then their relation is called linear correlation.

For example :

x	5	10	15	20	25
y	20	30	40	50	60

(B) Non-linear correlation : When the ratio of the changes between variables varies then relation between them is called nonlinear correlation. For example :

Expenditure on profit and advertisement

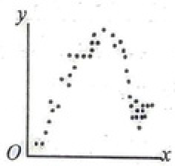


Fig. : Non-linear correlation

3. On basis of number of variables

(A) Simple correlation : Correlation between only two variates is called simple correlation and out of them one variable is said to be independent variable and other variable is dependent variable.

(B) Partial correlation : The inter relation between two variates only when the influence of other variate are kept constant is called partial correlation. For example : one taking amount of rainfall as constant, the study of correlation between average temperature and yield of wheat is called partial correlation.

(C) Multiple correlation : Study of joint effect of two or more independent variables on a variable is called multiple correlation.

For example :

Mathematical study of joint effect of amount of rainfall, nature of sand, average temperature on yield of wheat is called multiple correlation.

(D) Degree of correlation : Changes in two connected series may be either in same ratio or in varied ratio. This case can be measured by coefficient of correlation. The coefficient of correlation between two variables is expressed by r study the table given below for correlation order.

Value of r	Level of Correlation
$r = -1$	Perfect negative correlation
$-1 < r \leq -0.75$	High level of negative correlation
$-0.75 < r \leq -0.50$	Moderate level of negative correlation
$-0.50 < r < 0$	Low level of negative correlation
$r = 0$	Absence of correlation
$0 < r < 0.50$	Low level of positive correlation
$0.50 \leq r < 1$	High level positive correlation
$r = 1$	Perfect positive correlation

10.3 Methods of determining correlation

The following are main methods to find correlation :

(i) Scatter diagram or dot diagram.

(ii) Graphic method

(iii) Karl pearson's coefficient of correlation.

(iv) Spearman's ranking method.

(v) Concurrent deviation method.

(vi) Least square method.

In this chapter, we shall study only scatter diagram and karl person's coefficient of correlation.

1. Scatter diagram or dot diagram

It is an elementary method for finding the correlation between two variables. In this method, we plot on graph paper, the independent variable on the x-axis which denotes the abscissa and dependent variable on the y-axis which denotes the ordinate and plot in such a manner that each term of the two series, there is one point for one pair.

ESTIMATION OF THE RESULT OF CORRELATION BY SCATTER DIAGRAM

(i) POSITIVE CORRELATION

In a scatter diagram, if it appears that the direction of the points is from left bottom to the right above, then it shows positive correlation. If from these plotted points, we get a straight line from left to right (Fig. 1), then both the series have perfect positive correlation. If plotted points are not on the straight line but very near on both sides of it, then we have high level positive correlation (Fig. 2) and if they are scattered far away from the straight line, they have very low correlation (Fig. 3)

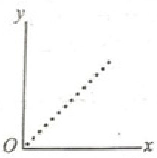


Fig. 1. Perfect positive correlation

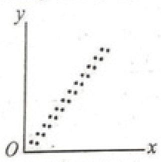


Fig. 2 : High level positive correlation

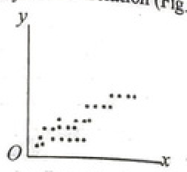


Fig. 3 : Low level Positive correlation

(ii) NEGATIVE CORRELATION

In a scatter diagram if it appears that the direction of the points is from the upper left to the right below, then it shows negative correlation. If they are on a straight line, then they have perfect negative correlation. If plotted points are not on the straight line but very near on both sides of it, then we have high level negative correlation and if they are scattered far away from the straight line, then they have low negative correlation.

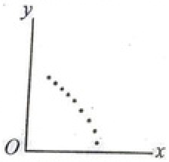


Fig. 4 : Perfect negative correlation

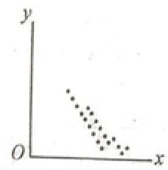


Fig. 5 : High level Negative correlation

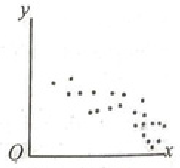
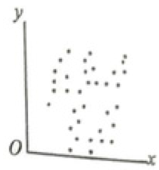


Fig. 6 : Low level Negative correlation

(iii) ABSENCE OF CORRELATION

If in the scatter diagram, the points are scattered all around and by these points there is no indication towards any direction, then there is an absence of correlation between two connected series.



Absence of Correlation

2. KARL PEARSON'S COEFFICIENT OF CORRELATION

To find numerical measure of direction and quantity of correlation, Karl Pearson gave a formula which is accurate from Maths point of view. This measure is based on arithmetic mean and standard deviation.

If pairs of two variable X and Y are $(x_1, y_1), (x_2, y_2).....(x_n, y_n)$ and mean are $\bar{x}$ and $\bar{y}$ respectively, then

$$\text{Karl Pearson's coefficient of correlation} = \frac{\text{covariance of } x \text{ and } y}{\sigma_x \, \sigma_y}$$

Where covariance of x and y

$$\text{cov}(x, y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

$$\sigma_x = \text{standard deviation of } X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\sigma_y = \text{standard deviation of } Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

n = number of paired observations

Karl Pearson's coefficient of correlation is represented by r.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

(i) For the calculation of r, the above formula should be used in those problems when $\bar{x}$ and $\bar{y}$ are integers.

(ii) For calculation of r, when $\bar{x}$ and $\bar{y}$ are not integer then coefficient of correlation for real values of x and y is find from the following formula.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$$

$$r = \frac{\Sigma x_i y_i - n\left(\frac{\Sigma x_i}{n}\right)\left(\frac{\Sigma y_i}{n}\right)}{\sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2} \times \sqrt{\frac{\Sigma y_i^2}{n} - \left(\frac{\Sigma y_i}{n}\right)^2}}$$

In easier form

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2} \times \sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$$

(iii) For the calculation of r, when $\bar{x}$ and $\bar{y}$ are not integers and variable values of x and y are bigger, then it is very difficult and time consuming to find the deviations and their squares. In such cases, it is convenient to use to following short-cut method. To find correlation from this method we use the following procedure :

1. First of all, we select assumed means 'a' and 'b' for both the series.
2. After that we find the deviations from the assumed means for both the series and get $u = x - a$ and $v = y - b$.
3. Now multiply the corresponding deviations of both series i.e. multiply u and v and find sum of their product $\Sigma u_i v_i$.
4. After that we find the sum of squares of deviations Σu_i^2 and Σv_i^2.
5. Finally we find coefficient of correlation by the following formula :

$$\frac{n\Sigma u_i v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{n\Sigma u_i^2 - (\Sigma u_i)^2} \times \sqrt{n\Sigma v_i^2 - (\Sigma v_i)^2}}$$

(iv) For Calculating r when class intervals of x's and y's values or difference in their values are equal then for convenience of calculation we ignore the class interval. In this, the coefficient of correlation is independent from change of origin and scale. Origin is shifted from assumed mean a of x and assumed mean b of y and taking new scale, class interval of x as h and class interval of y as k, the formula of step deviation method is

$$r_{xy} = r_{uv} = \frac{n\Sigma u_i v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{n\Sigma u_i^2 - (\Sigma u_i)^2} \times \sqrt{n\Sigma v_i^2 - (\Sigma v_i)^2}}$$

REGRESSION

A regression model is a mathematical equation that describes the relationship between two or more variables. It is also known as regression equation.

In regression analysis the *dependent variable* is one whose value is influenced or to be predicted. It is also known as regressed or explained variable while the variable which influences the values and is used for predicting the values of dependent variable is called as independent variable or regressor or predictor or explanatory variable.

Simple Regression : If we study the effect of a single independent variable on a dependent variable, it is called simple regression and such a model is known as simple regression model.

Multiple Regression : Studying the effect of two or more independent variables on a dependent variable is known as multiple regression and such a model is known as multiple regression model.

LINEAR REGRESSION

A regression equation, when plotted, may assume one or many possible shapes known as curve of regression. If this curve of regression is a straight line then it is said to be *line of regression* and such a regression is said to be *linear regression*. If the curve of regression is not a straight line then the regression is known as curvilinear regression or non linear regression.

Definition : A simple regression model that gives a straight line relationship between two variables is said to be linear regression model.

We always have two lines of regression in a simple linear regression model. Let the equation of the linear relationship between the two variables x and y be of the form y = a + bx, where y is treated as dependent variable and x is treated as independent variable. On treating these other way i.e. considering x as dependent variable and y as independent variable we can have the linear equation of the from x = c + dy; thus we have two lines of regression. The lines of regression give the best estimate to the values of one variable for any specific value of the other variable.

LINES OF REGRESSION

(i) Equation of line of regression of Y on X : If we choose the straight line in a linear regression model such that the sum of squares of deviations parallel to the axis of y is minimized, it is called the line of regression of Y on X. It gives the best estimates of Y for any given value of X.

Derivation

Let $y = a + bx$ be the line of regression of Y on X for the given data (x_i, y_i) i = 1, 2,....n.

Then for $$y = a + bx \qquad .....(1)$$

the normal equations are

$$\sum_{i=1}^{n} y = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx \Rightarrow \sum_{i=1}^{n} y = na + b\sum_{i=1}^{n} x \qquad .....(2)$$

and $$\sum_{i=1}^{n} xy = \sum_{i=1}^{n} ax + \sum_{i=1}^{n} bx^2 \Rightarrow \sum_{i=1}^{n} xy = a\sum_{i=1}^{n} x + b\sum_{i=1}^{n} x^2 \qquad ....(3)$$

Dividing equation (2) by n we have :

$$\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n} \Rightarrow \bar{y} = a + b\,\bar{x} \qquad .....(4)$$

as, $$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\Sigma x}{n} \text{ and } \bar{y} = \frac{y_1 + y_2 + ... + y_n}{n} = \frac{\Sigma y}{n}$$

Again dividing equation (3) by n we have

$$\frac{\Sigma xy}{n} = a\frac{\Sigma x}{n} + b\frac{\Sigma x^2}{n} \Rightarrow \frac{\Sigma xy}{n} = a\bar{x} + b\frac{\Sigma x^2}{n} \qquad .....(5)$$

Now we know that

$$\text{Cov}(x, y) = \mu_{11} = \frac{\Sigma xy}{n} - \bar{x}\,\bar{y} \Rightarrow \frac{\Sigma xy}{n} = \mu_{11} + \bar{x}\,\bar{y} \qquad ....(6)$$

Also $$\sigma_x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2 \Rightarrow \frac{\Sigma x^2}{n} = \sigma_x^2 + \bar{x}^2 \qquad ....(7)$$

Now equations (5) to (7) imply that

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b\left(\sigma_x^2 + \bar{x}^2\right)$$

$$\Rightarrow \quad \mu_{11} + \bar{x}\bar{y} = a\bar{x} + b\sigma_x^2 + b\bar{x}^2 = \bar{x}(a + b\bar{x}^2) + b\sigma_x^2$$

$$= \bar{x}\bar{y} + b\sigma_x^2$$

$$\Rightarrow \quad \mu_{11} = b\sigma_x^2$$

$$\Rightarrow \quad b = \frac{\mu_{11}}{\sigma_x^2} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x} \quad \left(\because r = \frac{\mu_{11}}{\sigma_x\sigma_y}\right)$$

From equation (4) and equation (7) it is clear that the required line passes through $(\bar{x}, \bar{y})$. Hence the equation of line passing through point $(\bar{x}, \bar{y})$ and having slope $b_{yx} = r\sigma_y/\sigma_x$ is

$$(y - \bar{y}) = b_{yx}\,(x - \bar{x})$$

$$\Rightarrow \quad y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad ....(8)$$

which is the required line of regression of Y on X.

(ii) Equation of line of regression of X on Y : In the linear regression model if the straight line is so chosen that the sum of squares of deviations parallel to axis of x are minimized, then it is called as line of regression of Y on X. Here x is treated as dependent variable and y is treated as independent variable.

It gives the best estimates of x for any given value of y.

It can be derived in the same manner as done in part (i) above, by interchanging the role of x and y. Its equation will be:

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad .....(9)$$

In case of perfect correlation (i.e. r = ±1) the equation of line of regression of y on x is

$$y - \bar{y} = \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

and equation of line of regression of x on y is

$$x-\bar{x} = \frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

both of which are similar.

Hence in general, we always have two lines of regression except in the case of perfect correlation ($r \pm 1$).

Note : 1. The line of regression of y on x as well as that of x on y both pass through $(\bar{x}, \bar{y})$. Hence $(\bar{x}, \bar{y})$ is the point of intersection of two lines of regression.

2. The equations for both the lines of regression are not reversible or interchangeable as basis and assumptions for deriving these equations are different.

3. If we have to predict the values of y for a given value of x then line of regression of y on x must be used, as in this case the predicted values will have minimum possible error

Properties of Regression Coefficient

(i) We know that $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ is regression coefficient of y on x and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ is regression coefficient of x on y, hence $b_{yx}.b_{xy} = r^2$.

$\Rightarrow$ $r = \pm\sqrt{b_{yx}b_{xy}}$ and the sign of r is same as that of the two regression coefficients.

(ii) We know that $r^2 \le 1 \Rightarrow b_{yx}\, b_{xy} \le 1$

$$\Rightarrow \quad b_{xy} \le \frac{1}{b_{yx}}$$

Now if $b_{yx} > 1 \Rightarrow b_{xy} < 1$. Hence if one of the regression coefficient is greater than unity, the other must be less than unity.

(iii) Arithmetic mean of regression coefficient is greater than the correlation coefficient (r), provided $r > 0$.

Arithmetic mean of regression coefficients

$$= \frac{1}{2}(b_{yx} + b_{xy}) = \frac{1}{2}\left(r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}\right)$$

Now $(\sigma_y - \sigma_x)^2 \ge 0 \Rightarrow \sigma_y^2 + \sigma_y^2 - 2\sigma_x\sigma_y \ge 0 \Rightarrow \frac{\sigma_x}{\sigma_y} + \frac{\sigma_y}{\sigma_x} \ge 2$

$$\Rightarrow \quad r\left(\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right) \ge 2r \qquad (\because r > 0)$$

$$\Rightarrow \quad \frac{1}{2}r\left(\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right) \ge r$$

(iv) Regression coefficients are independent of change of origin but not of scale.

ANGLE BETWEEN TWO LINES OF REGRESSION

Equation of line of regression of y on x is

$$y-\bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \text{ whose slope is } b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Equation of line of regression of x on y is :

$$x-\bar{x} = r\frac{\sigma_y}{\sigma_x}(y-\bar{y}) \Rightarrow y-\bar{y}\ \frac{\sigma_y}{r\sigma_x}(x-\bar{x})$$

$$\text{whose slope is } \frac{1}{b_{xy}} = \frac{\sigma_y}{r\sigma_x}$$

If θ is the angle between the two lines of regression then

$$\tan\theta = \frac{\frac{\sigma_y}{r\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \frac{r\sigma_y}{\sigma_x}\frac{\sigma_y}{r\sigma_x}} = \frac{(1-r^2)(\sigma_y/\sigma_x)}{\sigma_x^2+\sigma_y^2} \times \frac{\sigma_x^2}{r}$$

$$= \frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$$

$$\Rightarrow \quad \theta = \tan^{-1}\left\{\frac{1-r^2}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\}$$

As $\quad r^2 \le 1 \Rightarrow 1 - r^2 \ge 0 \le \theta < \pi/2$

Hence the acute angle (θ_1) between the two lines is

$$\tan\theta_1 = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Conversely $r^2 \le 1 \Rightarrow r^2 - 1 \le 0 \Rightarrow \pi/2 < Q \le \pi$, hence the obtuse angle (θ_2) between the two lines is

$$\tan\theta_2 = \left(\frac{r^2-1}{r}\right)\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

Hence we have the following possible cases :

(i) If $r = 0 \Rightarrow \tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Hence if the two variables are uncorrelated, the lines of regression become perpendicular to each other. Here as r = 0 the lines of regression are

$y - \bar{y} = 0 \Rightarrow y = \bar{y}$ and $x - \bar{x} = 0 \Rightarrow x = \bar{x}$

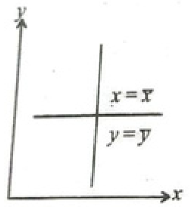


Fig. :

(ii) If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π.

This means that either the two lines are parallel to each other or the two lines coincide. But we know that both the lines intersect at the point $(\bar{x}, \bar{y})$, hence they cannot be parallel and must be coincident. Therefore in case of perfect correlation the two lines of regression are coincident.

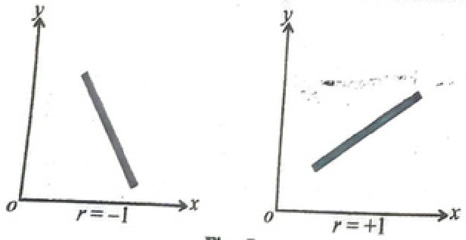
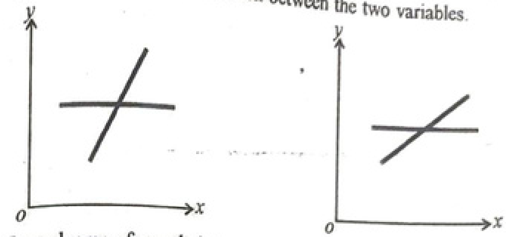


Fig. 7

Note : $r = 0 \Rightarrow \theta = \pi/2$ and $r = \pm 1 \Rightarrow \theta = 0$.

Hence we can conclude that for higher degree of correlation between the variables the angle between the lines is smaller, i.e., the two lines of regression are closer to each other and similarly we can say that larger angle between them indicates a poor degree of correlation between the variables. Thus by plotting the lines of regression on graph paper we can have a rough idea about the degree of correlation between the two variables.



Low degree of correlation

High degree of correlation

Fig. : 8

ILLUSTRATIVE EXAMPLES

Example 1 : Find the coefficient of correlation from following data :

x	2	3	5	7	3
y	15	17	4	5	4

Solution : Here $n = 5$

x	y	$x-\bar{x} = x-4$	$y-\bar{y} = y-9$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x}) \times (y-\bar{y})$
2	15	−2	6	4	36	−12
3	17	−1	8	1	64	−8
5	4	1	−5	1	25	−5
7	5	3	−4	9	16	−12
3	4	−1	−5	1	25	5
$\Sigma x_i = 20$	$\Sigma y_i = 45$			$\Sigma(x_i-\bar{x})^2 = 16$	$\Sigma(y_i-\bar{y})^2 = 166$	$\Sigma(x_i-\bar{x}) \times (y_i-\bar{y}) = 32$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{20}{5} = 4$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{45}{5} = 9$$

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}} = \frac{-32}{\sqrt{16} \times \sqrt{166}}$$

$$= \frac{-32}{4 \times 12.8841} = \frac{-32}{51.5364} = -0.62$$

■ **Example 2 : Find coefficient of correlation from the following data :**

x	1	2	3	4	5
y	2	5	7	8	10

Solution : The value of x and y are small numbers so by direct calculation method :

x	y	x^2	y^2	xy
1	2	1	4	2
2	5	4	25	10
3	7	9	49	21
4	8	16	64	32
5	10	25	100	50
$\Sigma x_i = 15$	Σy_1	$\Sigma x_1^2 = 55$	$\Sigma y_1^2 = 242$	$\Sigma x_i y_i = 115$

$$r = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_1)}{\sqrt{n\Sigma x_1^2 - (\Sigma x_1)^2} \times \sqrt{\Sigma y_1^2 - (\Sigma y_i)^2}}$$

$$r = \frac{5(1\cdot15) - 15(32)}{\sqrt{5(55) - (15)^2} \times \sqrt{5(242) - (32)^2}}$$

$$r = \frac{\sqrt{575 - 480}}{\sqrt{275 - 225} \times \sqrt{1210 - 1024}}$$

$$r = \frac{95}{\sqrt{50} \times \sqrt{186}} = 0.98$$

■ **Example 3. Find the coefficient of correlation from the following data :**

x	46	54	56	56	58	60	62	66
y	36	40	49	54	42	58	54	58

Solution : Assumed mean for x and y are $a = 56$ and $b = 49$ respectively.

x	y	$u = x - a$	$v = y - 6$	u^2	v^2	av
46	36	−10	−13	100	169	130
54	40	−2	−9	4	81	18
56	49	0	0	0	0	0
56	54	0	5	0	25	0
58	42	2	−7	4	49	−14
60	58	4	9	16	81	36
62	54	6	5	36	25	30
66	58	10	9	100	81	90
		$\Sigma u_i = 10$	$\Sigma v_i = -1$	$\Sigma u_i^2 = 260$	$\Sigma v_i^2 = 511$	$\Sigma u_i v_i = 290$

$$r = \frac{n\Sigma u_i v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{n\Sigma u_i^2 - (\Sigma u_i^2)} \times \sqrt{n\Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{8(290) - (10)(-1)}{\sqrt{8(260) - (10)^2} \times \sqrt{8(511) - (-1)^2}}$$

$$r = \frac{2320 + 10}{\sqrt{2080 - 100} \times \sqrt{4088 - 1}}$$

$$r = \frac{2330}{\sqrt{1980} \times \sqrt{4087}} = \frac{2330}{44.49 \times 63.92} = 0.81$$

■ **Example 4. Find the coefficient of correlation from the following data :**

x	155	165	175	185	195	205
y	77	62	52	52	47	42

Solution : Let $a = 175$ and $h = 10$ and $b = 52$, $k = 5$

x	y	$u=\frac{x-a}{h}$	$v=\frac{y-b}{k}$	u^2	v^2	uv
155	77	−2	5	4	25	−10
165	62	−1	2	1	4	−2
175	52	0	0	0	0	0
185	52	1	0	1	0	0
195	47	2	−1	4	1	−2
205	42	3	−2	9	4	−6
		$\Sigma u_1 = 3$	$\Sigma v_1 = 4$	$\Sigma u_1^2 = 19$	$\Sigma v_1^2 = 34$	$\Sigma u_1 v_1 = -20$

Here $n = 6$

$$r = \frac{n\Sigma u_i v_i - (\Sigma u_i)(\Sigma v_i)}{\sqrt{n\Sigma u_i^2 - (\Sigma u_i)^2} \times \sqrt{n\Sigma v_i^2 - (\Sigma v_i)^2}}$$

$$= \frac{6(-20) - (3)(4)}{\sqrt{6(19) - (3)^2} \times (34) - (4)^2}$$

$$= \frac{-120 - 12}{\sqrt{114 - 9} \times \sqrt{204 - 10}}$$

$$= \frac{-132}{\sqrt{105} \times \sqrt{188}}$$

$$= \frac{-132}{10.24 \times 13.71} = -0.94$$

Example 5. Calculate the coefficient of correlation and obtain the line of regression for the following data.

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	12	14	16	15

Obtain also an estimate for y which should correspond on an average to $x = 6.2$.

Solution : We can easily Solve and get

$$r_{xy} = r_{uv} = 0.95$$

As $u = x - 5$ and $v = y - 12$

$\Rightarrow$ $\bar{u} = \bar{x} - 5$ and $\bar{v} = \bar{y} - 12$

$$\sigma_u^2 = E\left[(u-\bar{u})^2\right] = E\left[\left[(x-5)-(\bar{x}-5)\right]^2\right] = E\left[(x-\bar{x})^2\right] = \sigma_x^2$$

and $$\sigma_v^2 = E\left[(v-\bar{v})^2\right] = E\left[\left[(y-12)-(\bar{y}-12)\right]^2\right] = E\left[(y-\bar{y})^2\right] = \sigma_y^2$$

Line of Regression of y on x is :

$$y - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$\Rightarrow \quad y - (\bar{v} + 12) = r_{uv}\frac{\sigma_v}{\sigma_u}[x - (\bar{u} + 5)]$$

$$\Rightarrow \quad y - (0 + 12) = 0.95 \times \frac{\sqrt{60/9}}{\sqrt{60/9}}[x - (0 + 5)]$$

$$\Rightarrow \quad y - 12 = 0.95(x - 5) \Rightarrow y = 0.95x + 7.25$$

Line of Regression of x on y is :

$$x - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$\Rightarrow \quad x - (\bar{u} + 5) = r_{uv}[y - (\bar{v} + 12)] \qquad (\sigma_x = \sigma_y \text{ as } \sigma_u = \sigma_v)$$

$$\Rightarrow \quad x - 5 = 0.95(y - 12)$$

$$\Rightarrow \quad x = 0.95y - 6.4$$

Since we need y at $x = 6.2$ hence line of regression y on x gives us the required estimate as :

$$y = 0.95x + 7.25$$

$$\Rightarrow \quad y = 0.95 \times 6.2 + 7.25$$

$$\Rightarrow \quad y = 13.14 \text{ at } x = 6.2$$

Example 6. In a partially destroyed laboratory on record of an analysis of correlation data, the following results only are legible,

Var $x = 9$, Regression equations : $8x - 10y + 66 = 0$, $40x - 18y = 214$

Find (i) The mean values of x and y.

(ii) The standard deviation of y.

(iii) The coefficient of correlation between x and y.

Solution : (i) We know that the mean value is the common point of intersection of the two lines of regression. Given regression equations are

$$8x - 10y + 66 = 0$$
$$40x - 18y = 214$$

Solving the above two equations we get $x = 13$ and $y = 17$

Hence the mean values are $\bar{x} = 13, \bar{y} = 17$

(ii) & (iii) First regression equation $\Rightarrow y = \frac{8}{10}x + \frac{66}{10}$

which can be treated as line of regression of y on x and second regression equation

$$\Rightarrow \quad x = \frac{18}{40}y + \frac{214}{40}$$

Which can be treated as line of regression of x on y

$$\Rightarrow \quad b_{yx} = \frac{8}{10} \text{ and } b_{xy} = \frac{18}{40}$$

As $$r^2 = b_{yx} \times b_{xy} = \frac{8}{10} \times \frac{18}{40} = \frac{9}{25}$$

$$= 0.36 \qquad \Rightarrow r = \pm 0.6$$

As both regression coefficients b_{yx} and b_{xy} are positive hence the correlation coefficient should also be positive and $r = 0.6$.

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{8}{10}$$

Here $r = 0.6, \sigma_x = \sqrt{9} = 3$ (given)

$$\Rightarrow \quad 0.6 \times \frac{\sigma_y}{3} = \frac{8}{10} \Rightarrow \sigma_y = \frac{4}{5} \times \frac{1}{0.2} = 4$$

Remark : If we take the first regression equation as line of regression of x on y i.e. $x = \frac{10}{8}y - \frac{66}{8}$ and the second regression equation as line of regression of y on

i.e., $$y = \frac{40}{18}x - \frac{214}{18} \text{ then } r^2 = \frac{10}{8} \times \frac{40}{18} = 2.778$$

$$\Rightarrow \quad r = 1.6$$

As $|r| > 1$ which is not possible, hence it is not permitted and the first regression equation should be line of regression of y on x.

■ **Example 7. Find Correlation coeff. between raimfall and temperature for following data :** **[R.U. 2016]**

Temperature	10	20	30	40	50	60	70
Rainfall	50	30	40	35	50	30	15

Solution :

x	y	u = x − a	v = y − b	u^2	v^2	uv
10	50	−30	15	900	225	−450
20	30	−20	−5	400	25	100
30	40	−10	5	100	25	−50
40	35	0	0	0	0	0
50	50	10	15	100	225	150
60	30	20	−5	400	25	−100
70	15	30	−20	900	400	−600
		$\Sigma u = 0$	$\Sigma v = 5$	$\Sigma u^2 = 2800$	$\Sigma v^2 = 925$	$\Sigma uv = -950$

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$\Rightarrow \quad r = \frac{7 \times -950 - 0}{\sqrt{7 \times 2800}\sqrt{7 \times 925 - 25}}$$

$$= \frac{-7 \times 950}{\sqrt{7 \times 400 \times 7 \times 6450}}$$

$$= \frac{-7 \times 950}{11243.66}$$

$$= -0.591$$

EXERCISES 10.1

1. Find the coefficient of correlation between series x and y from the following data :

	Series x	Series y
No of term	1000	1000
Standard deviation	4.5	3.6

Summation of product of deviation of x and y about their mean = 4800.

2. Find the coefficient of correlation between series x and y from the following data :

	series x	series y
No. of terms 8	8	
Sum of squares of deviation about mean	36	44

Summation of product of deviation of x and y about their mean = 24.

3. Find the Karl Pearson's coefficient of correlation from the following data :

x	−10	−5	0	5	10
y	5	9	7	11	13

4. Find the Karl Pearson's coefficient of correlation from the following data :

x	10	10	11	12	13	13	13	12	12	12
y	11	13	14	16	16	16	15	14	13	13

5. Find the coefficient of correlation between series x and y :

x	57	42	40	38	42	45	42	44	40	46	44	43
y	10	26	30	41	29	27	27	19	18	19	31	29

6. From the marks obtained by 10 student in Physics and Maths, compute the coefficient of correlation between the following marks :

Physics x	45	70	65	30	90	40	50	75	85	60
Maths y	35	90	70	40	95	40	65	80	80	50

7. The deviations of the series x and y form their respective assumed mean are following, calculate the coefficient of correlation by obtained data :

x	+5	−4	−2	+20	−10	0	+3	0	−15	−5
y	+5	−12	−7	+25	−10	−3	0	+2	−9	−15

Using change of origins and Scale, find Karl Pearson's coefficient of correlation from the following data :

x	21	28	42	56	63
y	90	100	130	160	170

9. From the following data obtain the two regression lines and the correlation coefficient. Also find the value of y when x = 82.

Sales (x)	100	98	78	85	110	93	80
Purchase (y)	85	90	70	72	95	81	74

10. Consider the two regression lines $3x + 2y = 26$ and $6x + y = 31$. (i) Find the mean values and correlation coefficient between x and y (ii) If the variance of y is 4, find the standard deviation of x.

11. Obtain the coefficient of correlation and lines of regression for the following data :

Age in years (x)	56	42	72	36	63	47	55	49	38	42	68	60
Blood pressure (y)	147	125	160	118	149	128	150	145	115	140	152	155

Also estimate the blood pressure when age = 45 years.

ANSWERS 10.1

1. +0.296 2. +0 − 6.3 3. +0.9 4. +0.78
5. − 0.73 6. +0.903 7. +0.89 8. +0.998
9. $y = 0.84x + 3.72$, $x = 1.12y + 1.28$; $r = 0.97$; $y = 72.6$
10. (i) $\bar{x} = 4$, $\bar{y} = 7$; $r = -0.5$ (ii) $\sigma_x = \frac{2}{3}$
11. $y = 1.138x + 80.778$; $y_{45} = 131.988$