



STATISTICS

❖ “Statistics may be rightly called the science of averages and their estimates.” – A.L.BOWLEY & A.L. BODDINGTON ❖

13.1 Introduction

We know that statistics deals with data collected for specific purposes. We can make decisions about the data by analysing and interpreting it. In earlier classes, we have studied methods of representing data graphically and in tabular form. This representation reveals certain salient features or characteristics of the data. We have also studied the methods of finding a representative value for the given data. This value is called the measure of central tendency. Recall mean (arithmetic mean), median and mode are three measures of central tendency. A *measure of central tendency* gives us a rough idea where data points are centred. But, in order to make better interpretation from the data, we should also have an idea how the data are scattered or how much they are bunched around a measure of central tendency.



Karl Pearson
(1857-1936)

Consider now the runs scored by two batsmen in their last ten matches as follows:

Batsman A : 30, 91, 0, 64, 42, 80, 30, 5, 117, 71

Batsman B : 53, 46, 48, 50, 53, 53, 58, 60, 57, 52

Clearly, the mean and median of the data are

| | Batsman A | Batsman B |
|--------|-----------|-----------|
| Mean | 53 | 53 |
| Median | 53 | 53 |

Recall that, we calculate the mean of a data (denoted by \bar{x}) by dividing the sum of the observations by the number of observations, i.e.,

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

Also, the median is obtained by first arranging the data in ascending or descending order and applying the following rule.

If the number of observations is odd, then the median is $\left(\frac{n+1}{2}\right)^{\text{th}}$ observation.

If the number of observations is even, then median is the mean of $\left(\frac{n}{2}\right)^{\text{th}}$ and

$\left(\frac{n}{2} + 1\right)^{\text{th}}$ observations.

We find that the mean and median of the runs scored by both the batsmen A and B are same i.e., 53. Can we say that the performance of two players is same? Clearly No, because the variability in the scores of batsman A is from 0 (minimum) to 117 (maximum). Whereas, the range of the runs scored by batsman B is from 46 to 60.

Let us now plot the above scores as dots on a number line. We find the following diagrams:

For batsman A



Fig 13.1

For batsman B

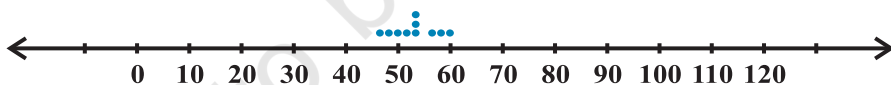


Fig 13.2

We can see that the dots corresponding to batsman B are close to each other and are clustering around the measure of central tendency (mean and median), while those corresponding to batsman A are scattered or more spread out.

Thus, the measures of central tendency are not sufficient to give complete information about a given data. Variability is another factor which is required to be studied under statistics. Like '*measures of central tendency*' we want to have a single number to describe variability. This single number is called a '*measure of dispersion*'. In this Chapter, we shall learn some of the important measures of dispersion and their methods of calculation for ungrouped and grouped data.

13.2 Measures of Dispersion

The dispersion or scatter in a data is measured on the basis of the observations and the types of the measure of central tendency, used there. There are following measures of dispersion:

(i) Range, (ii) Quartile deviation, (iii) Mean deviation, (iv) Standard deviation.

In this Chapter, we shall study all of these measures of dispersion except the quartile deviation.

13.3 Range

Recall that, in the example of runs scored by two batsmen A and B, we had some idea of variability in the scores on the basis of minimum and maximum runs in each series. To obtain a single number for this, we find the difference of maximum and minimum values of each series. This difference is called the 'Range' of the data.

In case of batsman A, $\text{Range} = 117 - 0 = 117$ and for batsman B, $\text{Range} = 60 - 46 = 14$. Clearly, $\text{Range of A} > \text{Range of B}$. Therefore, the scores are scattered or dispersed in case of A while for B these are close to each other.

Thus, $\text{Range of a series} = \text{Maximum value} - \text{Minimum value}$.

The range of data gives us a rough idea of variability or scatter but does not tell about the dispersion of the data from a measure of central tendency. For this purpose, we need some other measure of variability. Clearly, such measure must depend upon the difference (or deviation) of the values from the central tendency.

The important measures of dispersion, which depend upon the deviations of the observations from a central tendency are mean deviation and standard deviation. Let us discuss them in detail.

13.4 Mean Deviation

Recall that the deviation of an observation x from a fixed value ' a ' is the difference $x - a$. In order to find the dispersion of values of x from a central value ' a ', we find the deviations about a . An absolute measure of dispersion is the mean of these deviations. To find the mean, we must obtain the sum of the deviations. But, we know that a measure of central tendency lies between the maximum and the minimum values of the set of observations. Therefore, some of the deviations will be negative and some positive. Thus, the sum of deviations may vanish. Moreover, the sum of the deviations from mean (\bar{x}) is zero.

$$\text{Also} \quad \text{Mean of deviations} = \frac{\text{Sum of deviations}}{\text{Number of observations}} = \frac{0}{n} = 0$$

Thus, finding the mean of deviations about mean is not of any use for us, as far as the measure of dispersion is concerned.

Remember that, in finding a suitable measure of dispersion, we require the distance of each value from a central tendency or a fixed number 'a'. Recall, that the absolute value of the difference of two numbers gives the distance between the numbers when represented on a number line. Thus, to find the measure of dispersion from a fixed number 'a' we may take the mean of the absolute values of the deviations from the central value. This mean is called the '*mean deviation*'. Thus mean deviation about a central value 'a' is the mean of the absolute values of the deviations of the observations from 'a'. The mean deviation from 'a' is denoted as M.D. (a). Therefore,

$$\text{M.D.}(a) = \frac{\text{Sum of absolute values of deviations from 'a'}}{\text{Number of observations}}.$$

Remark Mean deviation may be obtained from any measure of central tendency. However, mean deviation from mean and median are commonly used in statistical studies.

Let us now learn how to calculate mean deviation about mean and mean deviation about median for various types of data

13.4.1 Mean deviation for ungrouped data Let n observations be $x_1, x_2, x_3, \dots, x_n$. The following steps are involved in the calculation of mean deviation about mean or median:

Step 1 Calculate the measure of central tendency about which we are to find the mean deviation. Let it be 'a'.

Step 2 Find the deviation of each x_i from a , i.e., $x_1 - a, x_2 - a, x_3 - a, \dots, x_n - a$


Step 3 Find the absolute values of the deviations, i.e., drop the minus sign (-), if it is there, i.e., $|x_1 - a|, |x_2 - a|, |x_3 - a|, \dots, |x_n - a|$

Step 4 Find the mean of the absolute values of the deviations. This mean is the mean deviation about a , i.e.,

$$\text{M.D.}(a) = \frac{\sum_{i=1}^n |x_i - a|}{n}$$

Thus $\text{M.D.}(\bar{x}) = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$, where \bar{x} = Mean

and $\text{M.D.}(M) = \frac{1}{n} \sum_{i=1}^n |x_i - M|$, where M = Median

 **Note** In this Chapter, we shall use the symbol M to denote median unless stated otherwise. Let us now illustrate the steps of the above method in following examples.

Example 1 Find the mean deviation about the mean for the following data:

6, 7, 10, 12, 13, 4, 8, 12

Solution We proceed step-wise and get the following:

Step 1 Mean of the given data is

$$\bar{x} = \frac{6+7+10+12+13+4+8+12}{8} = \frac{72}{8} = 9$$

Step 2 The deviations of the respective observations from the mean \bar{x} , i.e., $x_i - \bar{x}$ are

6 – 9, 7 – 9, 10 – 9, 12 – 9, 13 – 9, 4 – 9, 8 – 9, 12 – 9,


or –3, –2, 1, 3, 4, –5, –1, 3

Step 3 The absolute values of the deviations, i.e., $|x_i - \bar{x}|$ are

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

Step 4 The required mean deviation about the mean is

$$\begin{aligned} \text{M.D. } (\bar{x}) &= \frac{\sum_{i=1}^8 |x_i - \bar{x}|}{8} \\ &= \frac{3+2+1+3+4+5+1+3}{8} = \frac{22}{8} = 2.75 \end{aligned}$$

 **Note** Instead of carrying out the steps every time, we can carry on calculation, step-wise without referring to steps.

Example 2 Find the mean deviation about the mean for the following data :

12, 3, 18, 17, 4, 9, 17, 19, 20, 15, 8, 17, 2, 3, 16, 11, 3, 1, 0, 5

Solution We have to first find the mean (\bar{x}) of the given data

$$\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{200}{20} = 10$$

The respective absolute values of the deviations from mean, i.e., $|x_i - \bar{x}|$ are

2, 7, 8, 7, 6, 1, 7, 9, 10, 5, 2, 7, 8, 7, 6, 1, 7, 9, 10, 5

Therefore
$$\sum_{i=1}^{20} |x_i - \bar{x}| = 124$$

and
$$\text{M.D. } (\bar{x}) = \frac{124}{20} = 6.2$$

Example 3 Find the mean deviation about the median for the following data:

3, 9, 5, 3, 12, 10, 18, 4, 7, 19, 21.

Solution Here the number of observations is 11 which is odd. Arranging the data into ascending order, we have 3, 3, 4, 5, 7, 9, 10, 12, 18, 19, 21

Now
$$\text{Median} = \left(\frac{11 + 1}{2} \right)^{\text{th}} \text{ or } 6^{\text{th}} \text{ observation} = 9$$

The absolute values of the respective deviations from the median, i.e., $|x_i - M|$ are

6, 6, 5, 4, 2, 0, 1, 3, 9, 10, 12

Therefore
$$\sum_{i=1}^{11} |x_i - M| = 58$$

and
$$\text{M.D. } (M) = \frac{1}{11} \sum_{i=1}^{11} |x_i - M| = \frac{1}{11} \times 58 = 5.27$$

13.4.2 Mean deviation for grouped data We know that data can be grouped into two ways :

- (a) Discrete frequency distribution,
- (b) Continuous frequency distribution.

Let us discuss the method of finding mean deviation for both types of the data.

(a) Discrete frequency distribution Let the given data consist of n distinct values x_1, x_2, \dots, x_n occurring with frequencies f_1, f_2, \dots, f_n respectively. This data can be represented in the tabular form as given below, and is called *discrete frequency distribution*:

$$\begin{array}{ccccccc} x : & x_1 & & x_2 & & x_3 & \dots & x_n \\ f : & f_1 & & f_2 & & f_3 & \dots & f_n \end{array}$$

(i) Mean deviation about mean

First of all we find the mean \bar{x} of the given data by using the formula

$$\bar{x} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i} = \frac{1}{N} \sum_{i=1}^n x_i f_i,$$

where $\sum_{i=1}^n x_i f_i$ denotes the sum of the products of observations x_i with their respective

frequencies f_i and $N = \sum_{i=1}^n f_i$ is the sum of the frequencies.

Then, we find the deviations of observations x_i from the mean \bar{x} and take their absolute values, i.e., $|x_i - \bar{x}|$ for all $i = 1, 2, \dots, n$.

After this, find the mean of the absolute values of the deviations, which is the required mean deviation about the mean. Thus

$$\text{M.D.}(\bar{x}) = \frac{\sum_{i=1}^n f_i |x_i - \bar{x}|}{\sum_{i=1}^n f_i} = \frac{1}{N} \sum_{i=1}^n f_i |x_i - \bar{x}|$$

(ii) Mean deviation about median To find mean deviation about median, we find the median of the given discrete frequency distribution. For this the observations are arranged in ascending order. After this the cumulative frequencies are obtained. Then, we identify

the observation whose cumulative frequency is equal to or just greater than $\frac{N}{2}$, where

N is the sum of frequencies. This value of the observation lies in the middle of the data, therefore, it is the required median. After finding median, we obtain the mean of the absolute values of the deviations from median. Thus,

$$\text{M.D.}(M) = \frac{1}{N} \sum_{i=1}^n f_i |x_i - M|$$

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

| | | | | | | |
|-------|---|---|----|---|----|----|
| x_i | 2 | 5 | 6 | 8 | 10 | 12 |
| f_i | 2 | 8 | 10 | 7 | 8 | 5 |

Solution Let us make a Table 13.1 of the given data and append other columns after calculations.

Table 13.1

| x_i | f_i | $f_i x_i$ | $ x_i - \bar{x} $ | $f_i x_i - \bar{x} $ |
|-------|-------|-----------|-------------------|-----------------------|
| 2 | 2 | 4 | 5.5 | 11 |
| 5 | 8 | 40 | 2.5 | 20 |
| 6 | 10 | 60 | 1.5 | 15 |
| 8 | 7 | 56 | 0.5 | 3.5 |
| 10 | 8 | 80 | 2.5 | 20 |
| 12 | 5 | 60 | 4.5 | 22.5 |
| | 40 | 300 | | 92 |

$$N = \sum_{i=1}^6 f_i = 40, \quad \sum_{i=1}^6 f_i x_i = 300, \quad \sum_{i=1}^6 f_i |x_i - \bar{x}| = 92$$

Therefore
$$\bar{x} = \frac{1}{N} \sum_{i=1}^6 f_i x_i = \frac{1}{40} \times 300 = 7.5$$

and
$$\text{M. D. } (\bar{x}) = \frac{1}{N} \sum_{i=1}^6 f_i |x_i - \bar{x}| = \frac{1}{40} \times 92 = 2.3$$

Example 5 Find the mean deviation about the median for the following data:

| | | | | | | | | |
|-------|---|---|---|----|----|----|----|----|
| x_i | 3 | 6 | 9 | 12 | 13 | 15 | 21 | 22 |
| f_i | 3 | 4 | 5 | 2 | 4 | 5 | 4 | 3 |

Solution The given observations are already in ascending order. Adding a row corresponding to cumulative frequencies to the given data, we get (Table 13.2).

Table 13.2

| | | | | | | | | |
|--------|---|---|----|----|----|----|----|----|
| x_i | 3 | 6 | 9 | 12 | 13 | 15 | 21 | 22 |
| f_i | 3 | 4 | 5 | 2 | 4 | 5 | 4 | 3 |
| $c.f.$ | 3 | 7 | 12 | 14 | 18 | 23 | 27 | 30 |

Now, $N=30$ which is even.

Median is the mean of the 15th and 16th observations. Both of these observations lie in the cumulative frequency 18, for which the corresponding observation is 13.

Therefore, Median $M = \frac{15^{\text{th}} \text{ observation} + 16^{\text{th}} \text{ observation}}{2} = \frac{13 + 13}{2} = 13$

Now, absolute values of the deviations from median, i.e., $|x_i - M|$ are shown in Table 13.3.

Table 13.3

| | | | | | | | | |
|-----------------|----|----|----|---|---|----|----|----|
| $ x_i - M $ | 10 | 7 | 4 | 1 | 0 | 2 | 8 | 9 |
| f_i | 3 | 4 | 5 | 2 | 4 | 5 | 4 | 3 |
| $f_i x_i - M $ | 30 | 28 | 20 | 2 | 0 | 10 | 32 | 27 |

We have $\sum_{i=1}^8 f_i = 30$ and $\sum_{i=1}^8 f_i |x_i - M| = 149$

Therefore $M. D. (M) = \frac{1}{N} \sum_{i=1}^8 f_i |x_i - M|$
 $= \frac{1}{30} \times 149 = 4.97.$

(b) Continuous frequency distribution A continuous frequency distribution is a series in which the data are classified into different class-intervals without gaps along with their respective frequencies.

For example, marks obtained by 100 students are presented in a continuous frequency distribution as follows :

| | | | | | | |
|--------------------|------|-------|-------|-------|-------|-------|
| Marks obtained | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
| Number of Students | 12 | 18 | 27 | 20 | 17 | 6 |

(i) Mean deviation about mean While calculating the mean of a continuous frequency distribution, we had made the assumption that the frequency in each class is centred at its mid-point. Here also, we write the mid-point of each given class and proceed further as for a discrete frequency distribution to find the mean deviation.

Let us take the following example.

Example 6 Find the mean deviation about the mean for the following data.

| | | | | | | | |
|--------------------|-------|-------|-------|-------|-------|-------|-------|
| Marks obtained | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
| Number of students | 2 | 3 | 8 | 14 | 8 | 3 | 2 |

Solution We make the following Table 13.4 from the given data :

Table 13.4

| Marks obtained | Number of students f_i | Mid-points x_i | $f_i x_i$ | $ x_i - \bar{x} $ | $f_i x_i - \bar{x} $ |
|----------------|-----------------------------|---------------------|-----------|-------------------|-----------------------|
| 10-20 | 2 | 15 | 30 | 30 | 60 |
| 20-30 | 3 | 25 | 75 | 20 | 60 |
| 30-40 | 8 | 35 | 280 | 10 | 80 |
| 40-50 | 14 | 45 | 630 | 0 | 0 |
| 50-60 | 8 | 55 | 440 | 10 | 80 |
| 60-70 | 3 | 65 | 195 | 20 | 60 |
| 70-80 | 2 | 75 | 150 | 30 | 60 |
| | 40 | | 1800 | | 400 |

Here $N = \sum_{i=1}^7 f_i = 40$, $\sum_{i=1}^7 f_i x_i = 1800$, $\sum_{i=1}^7 f_i |x_i - \bar{x}| = 400$

Therefore $\bar{x} = \frac{1}{N} \sum_{i=1}^7 f_i x_i = \frac{1800}{40} = 45$

and $M.D.(\bar{x}) = \frac{1}{N} \sum_{i=1}^7 f_i |x_i - \bar{x}| = \frac{1}{40} \times 400 = 10$

Shortcut method for calculating mean deviation about mean We can avoid the tedious calculations of computing \bar{x} by following step-deviation method. Recall that in this method, we take an assumed mean which is in the middle or just close to it in the data. Then deviations of the observations (or mid-points of classes) are taken from the

assumed mean. This is nothing but the shifting of origin from zero to the assumed mean on the number line, as shown in Fig 13.3

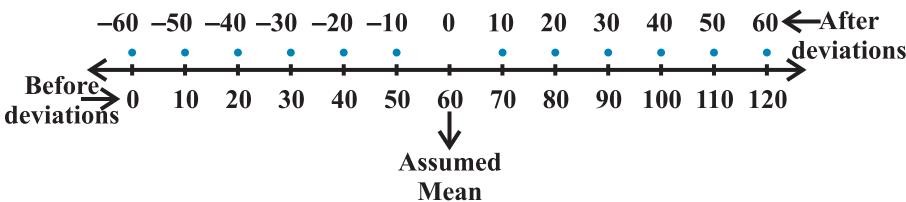


Fig 13.3

If there is a common factor of all the deviations, we divide them by this common factor to further simplify the deviations. These are known as step-deviations. The process of taking step-deviations is the change of scale on the number line as shown in Fig 13.4

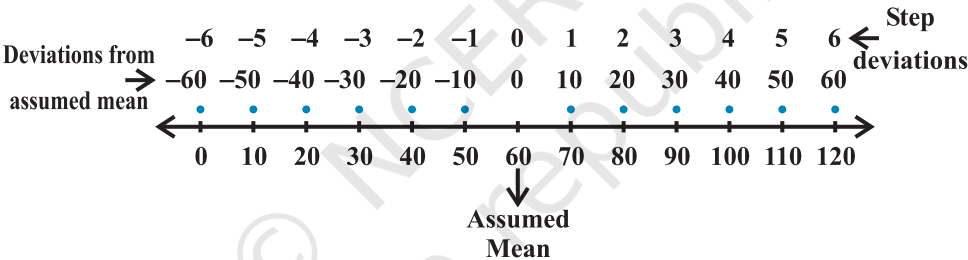


Fig 13.4

The deviations and step-deviations reduce the size of the observations, so that the computations viz. multiplication, etc., become simpler. Let, the new variable be denoted

by $d_i = \frac{x_i - a}{h}$, where 'a' is the assumed mean and h is the common factor. Then, the mean \bar{x} by step-deviation method is given by

$$\bar{x} = a + \frac{\sum_{i=1}^n f_i d_i}{N} \times h$$

Let us take the data of Example 6 and find the mean deviation by using step-deviation method.

Take the assumed mean $a = 45$ and $h = 10$, and form the following Table 13.5.

Table 13.5


| Marks obtained | Number of students | Mid-points | $d_i = \frac{x_i - 45}{10}$ | $f_i d_i$ | $ x_i - \bar{x} $ | $f_i x_i - \bar{x} $ |
|----------------|--------------------|------------|-----------------------------|-----------|-------------------|-----------------------|
| | f_i | x_i | | | | |
| 10-20 | 2 | 15 | -3 | -6 | 30 | 60 |
| 20-30 | 3 | 25 | -2 | -6 | 20 | 60 |
| 30-40 | 8 | 35 | -1 | -8 | 10 | 80 |
| 40-50 | 14 | 45 | 0 | 0 | 0 | 0 |
| 50-60 | 8 | 55 | 1 | 8 | 10 | 80 |
| 60-70 | 3 | 65 | 2 | 6 | 20 | 60 |
| 70-80 | 2 | 75 | 3 | 6 | 30 | 60 |
| | 40 | | | 0 | | 400 |

Therefore

$$\begin{aligned}\bar{x} &= a + \frac{\sum_{i=1}^7 f_i d_i}{N} \times h \\ &= 45 + \frac{0}{40} \times 10 = 45\end{aligned}$$

and

$$\text{M.D. } (\bar{x}) = \frac{1}{N} \sum_{i=1}^7 f_i |x_i - \bar{x}| = \frac{400}{40} = 10$$

 **Note** The step deviation method is applied to compute \bar{x} . Rest of the procedure is same.

(ii) Mean deviation about median The process of finding the mean deviation about median for a continuous frequency distribution is similar as we did for mean deviation about the mean. The only difference lies in the replacement of the mean by median while taking deviations.

Let us recall the process of finding median for a continuous frequency distribution.

The data is first arranged in ascending order. Then, the median of continuous frequency distribution is obtained by first identifying the class in which median lies (median class) and then applying the formula

$$\text{Median} = l + \frac{\frac{N}{2} - C}{f} \times h$$

where median class is the class interval whose cumulative frequency is just greater than or equal to $\frac{N}{2}$, N is the sum of frequencies, l, f, h and C are, respectively the lower limit, the frequency, the width of the median class and C the cumulative frequency of the class just preceding the median class. After finding the median, the absolute values of the deviations of mid-point x_i of each class from the median i.e., $|x_i - M|$ are obtained.

Then
$$\text{M.D. (M)} = \frac{1}{N} \sum_{i=1}^n f_i |x_i - M|$$

The process is illustrated in the following example:

Example 7 Calculate the mean deviation about median for the following data :

| | | | | | | |
|-----------|------|-------|-------|-------|-------|-------|
| Class | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
| Frequency | 6 | 7 | 15 | 16 | 4 | 2 |

Solution Form the following Table 13.6 from the given data :

Table 13.6

| Class | Frequency | Cumulative frequency | Mid-points | $ x_i - \text{Med.} $ | $f_i x_i - \text{Med.} $ |
|-------|-----------|----------------------|------------|-----------------------|---------------------------|
| | f_i | (c.f.) | x_i | | |
| 0-10 | 6 | 6 | 5 | 23 | 138 |
| 10-20 | 7 | 13 | 15 | 13 | 91 |
| 20-30 | 15 | 28 | 25 | 3 | 45 |
| 30-40 | 16 | 44 | 35 | 7 | 112 |
| 40-50 | 4 | 48 | 45 | 17 | 68 |
| 50-60 | 2 | 50 | 55 | 27 | 54 |
| | 50 | | | | 508 |

The class interval containing $\frac{N}{2}$ or 25th item is 20-30. Therefore, 20-30 is the median class. We know that

$$\text{Median} = l + \frac{\frac{N}{2} - C}{f} \times h$$

Here $l = 20$, $C = 13$, $f = 15$, $h = 10$ and $N = 50$

Therefore, $\text{Median} = 20 + \frac{25 - 13}{15} \times 10 = 20 + 8 = 28$

Thus, Mean deviation about median is given by

$$\text{M.D. (M)} = \frac{1}{N} \sum_{i=1}^6 f_i |x_i - M| = \frac{1}{50} \times 508 = 10.16$$

EXERCISE 13.1

Find the mean deviation about the mean for the data in Exercises 1 and 2.

1. 4, 7, 8, 9, 10, 12, 13, 17
2. 38, 70, 48, 40, 42, 55, 63, 46, 54, 44

Find the mean deviation about the median for the data in Exercises 3 and 4.

3. 13, 17, 16, 14, 11, 13, 10, 16, 11, 18, 12, 17
4. 36, 72, 46, 42, 60, 45, 53, 46, 51, 49

Find the mean deviation about the mean for the data in Exercises 5 and 6.

5.

| | | | | | |
|-------|---|----|----|----|----|
| x_i | 5 | 10 | 15 | 20 | 25 |
| f_i | 7 | 4 | 6 | 3 | 5 |
6.

| | | | | | |
|-------|----|----|----|----|----|
| x_i | 10 | 30 | 50 | 70 | 90 |
| f_i | 4 | 24 | 28 | 16 | 8 |

Find the mean deviation about the median for the data in Exercises 7 and 8.

7.

| | | | | | | |
|-------|---|---|---|----|----|----|
| x_i | 5 | 7 | 9 | 10 | 12 | 15 |
| f_i | 8 | 6 | 2 | 2 | 2 | 6 |
8.

| | | | | | |
|-------|----|----|----|----|----|
| x_i | 15 | 21 | 27 | 30 | 35 |
| f_i | 3 | 5 | 6 | 7 | 8 |

Find the mean deviation about the mean for the data in Exercises 9 and 10.

| | | | | | | | | | |
|----|---------------------|-------|---------|---------|---------|---------|---------|---------|---------|
| 9. | Income per day in ₹ | 0-100 | 100-200 | 200-300 | 300-400 | 400-500 | 500-600 | 600-700 | 700-800 |
| | Number of persons | 4 | 8 | 9 | 10 | 7 | 5 | 4 | 3 |

| | | | | | | | |
|-----|----------------|--------|---------|---------|---------|---------|---------|
| 10. | Height in cms | 95-105 | 105-115 | 115-125 | 125-135 | 135-145 | 145-155 |
| | Number of boys | 9 | 13 | 26 | 30 | 12 | 10 |

11. Find the mean deviation about median for the following data :

| | | | | | | |
|-----------------|------|-------|-------|-------|-------|-------|
| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
| Number of Girls | 6 | 8 | 14 | 16 | 4 | 2 |

12. Calculate the mean deviation about median age for the age distribution of 100 persons given below:

| | | | | | | | | |
|----------------|-------|-------|-------|-------|-------|-------|-------|-------|
| Age (in years) | 16-20 | 21-25 | 26-30 | 31-35 | 36-40 | 41-45 | 46-50 | 51-55 |
| Number | 5 | 6 | 12 | 14 | 26 | 12 | 16 | 9 |

[**Hint** Convert the given data into continuous frequency distribution by subtracting 0.5 from the lower limit and adding 0.5 to the upper limit of each class interval]

13.4.3 Limitations of mean deviation In a series, where the degree of variability is very high, the median is not a representative central tendency. Thus, the mean deviation about median calculated for such series can not be fully relied.

The sum of the deviations from the mean (minus signs ignored) is more than the sum of the deviations from median. Therefore, the mean deviation about the mean is not very scientific. Thus, in many cases, mean deviation may give unsatisfactory results. Also mean deviation is calculated on the basis of absolute values of the deviations and therefore, cannot be subjected to further algebraic treatment. This implies that we must have some other measure of dispersion. Standard deviation is such a measure of dispersion.

13.5 Variance and Standard Deviation

Recall that while calculating mean deviation about mean or median, the absolute values of the deviations were taken. The absolute values were taken to give meaning to the mean deviation, otherwise the deviations may cancel among themselves.

Another way to overcome this difficulty which arose due to the signs of deviations, is to take squares of all the deviations. Obviously all these squares of deviations are

non-negative. Let $x_1, x_2, x_3, \dots, x_n$ be n observations and \bar{x} be their mean. Then

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

If this sum is zero, then each $(x_i - \bar{x})$ has to be zero. This implies that there is no dispersion at all as all observations are equal to the mean \bar{x} .

If $\sum_{i=1}^n (x_i - \bar{x})^2$ is small, this indicates that the observations $x_1, x_2, x_3, \dots, x_n$ are close to the mean \bar{x} and therefore, there is a lower degree of dispersion. On the contrary, if this sum is large, there is a higher degree of dispersion of the observations from the mean \bar{x} . Can we thus say that the sum $\sum_{i=1}^n (x_i - \bar{x})^2$ is a reasonable indicator of the degree of dispersion or scatter?

Let us take the set A of six observations 5, 15, 25, 35, 45, 55. The mean of the observations is $\bar{x} = 30$. The sum of squares of deviations from \bar{x} for this set is

$$\begin{aligned} \sum_{i=1}^6 (x_i - \bar{x})^2 &= (5-30)^2 + (15-30)^2 + (25-30)^2 + (35-30)^2 + (45-30)^2 + (55-30)^2 \\ &= 625 + 225 + 25 + 25 + 225 + 625 = 1750 \end{aligned}$$

Let us now take another set B of 31 observations 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45. The mean of these observations is $\bar{y} = 30$

Note that both the sets A and B of observations have a mean of 30.

Now, the sum of squares of deviations of observations for set B from the mean \bar{y} is given by

$$\begin{aligned} \sum_{i=1}^{31} (y_i - \bar{y})^2 &= (15-30)^2 + (16-30)^2 + (17-30)^2 + \dots + (44-30)^2 + (45-30)^2 \\ &= (-15)^2 + (-14)^2 + \dots + (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 + \dots + 14^2 + 15^2 \\ &= 2 [15^2 + 14^2 + \dots + 1^2] \\ &= 2 \times \frac{15 \times (15+1) (30+1)}{6} = 5 \times 16 \times 31 = 2480 \end{aligned}$$

(Because sum of squares of first n natural numbers = $\frac{n(n+1)(2n+1)}{6}$. Here $n = 15$)

If $\sum_{i=1}^n (x_i - \bar{x})^2$ is simply our measure of dispersion or scatter about mean, we

will tend to say that the set A of six observations has a lesser dispersion about the mean than the set B of 31 observations, even though the observations in set A are more scattered from the mean (the range of deviations being from -25 to 25) than in the set B (where the range of deviations is from -15 to 15).

This is also clear from the following diagrams.

For the set A, we have

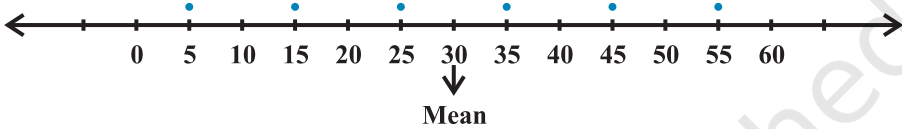


Fig 13.5

For the set B, we have

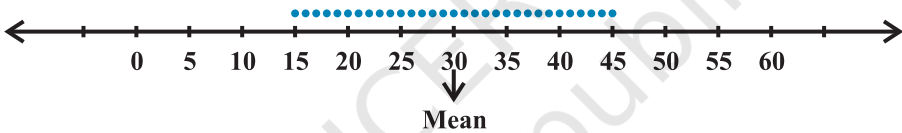


Fig 13.6

Thus, we can say that the sum of squares of deviations from the mean is not a proper measure of dispersion. To overcome this difficulty we take the mean of the squares of

the deviations, i.e., we take $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. In case of the set A, we have

$$\text{Mean} = \frac{1}{6} \times 1750 = 291.67 \text{ and in case of the set B, it is } \frac{1}{31} \times 2480 = 80.$$

This indicates that the scatter or dispersion is more in set A than the scatter or dispersion in set B, which confirms with the geometrical representation of the two sets.

Thus, we can take $\frac{1}{n} \sum (x_i - \bar{x})^2$ as a quantity which leads to a proper measure

of dispersion. This number, i.e., mean of the squares of the deviations from mean is called the **variance** and is denoted by σ^2 (read as sigma square). Therefore, the variance of n observations x_1, x_2, \dots, x_n is given by

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

13.5.1 Standard Deviation In the calculation of variance, we find that the units of individual observations x_i and the unit of their mean \bar{x} are different from that of variance, since variance involves the sum of squares of $(x_i - \bar{x})$. For this reason, the proper measure of dispersion about the mean of a set of observations is expressed as positive square-root of the variance and is called *standard deviation*. Therefore, the standard deviation, usually denoted by σ , is given by

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

Let us take the following example to illustrate the calculation of variance and hence, standard deviation of ungrouped data.

Example 8 Find the variance of the following data:

6, 8, 10, 12, 14, 16, 18, 20, 22, 24

Solution From the given data we can form the following Table 13.7. The mean is calculated by step-deviation method taking 14 as assumed mean. The number of observations is $n = 10$

Table 13.7

| x_i | $d_i = \frac{x_i - 14}{2}$ | Deviations from mean $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ |
|-------|----------------------------|---|---------------------|
| 6 | -4 | -9 | 81 |
| 8 | -3 | -7 | 49 |
| 10 | -2 | -5 | 25 |
| 12 | -1 | -3 | 9 |
| 14 | 0 | -1 | 1 |
| 16 | 1 | 1 | 1 |
| 18 | 2 | 3 | 9 |
| 20 | 3 | 5 | 25 |
| 22 | 4 | 7 | 49 |
| 24 | 5 | 9 | 81 |
| | 5 | | 330 |

Therefore $\text{Mean } \bar{x} = \text{assumed mean} + \frac{\sum_{i=1}^n d_i}{n} \times h = 14 + \frac{5}{10} \times 2 = 15$

and $\text{Variance } (\sigma^2) = \frac{1}{n} \sum_{i=1}^{10} (x_i - \bar{x})^2 = \frac{1}{10} \times 330 = 33$

Thus $\text{Standard deviation } (\sigma) = \sqrt{33} = 5.74$

13.5.2 Standard deviation of a discrete frequency distribution Let the given discrete frequency distribution be

$$\begin{array}{l} x: \quad x_1, \quad x_2, \quad x_3, \dots, x_n \\ f: \quad f_1, \quad f_2, \quad f_3, \dots, f_n \end{array}$$

In this case standard deviation $(\sigma) = \sqrt{\frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^2} \dots (2)$

where $N = \sum_{i=1}^n f_i$.

Let us take up following example.

Example 9 Find the variance and standard deviation for the following data:

| | | | | | | | |
|-------|---|---|----|----|----|----|----|
| x_i | 4 | 8 | 11 | 17 | 20 | 24 | 32 |
| f_i | 3 | 5 | 9 | 5 | 4 | 3 | 1 |

Solution Presenting the data in tabular form (Table 13.8), we get

Table 13.8

| x_i | f_i | $f_i x_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $f_i (x_i - \bar{x})^2$ |
|-------|-------|-----------|-----------------|---------------------|-------------------------|
| 4 | 3 | 12 | -10 | 100 | 300 |
| 8 | 5 | 40 | -6 | 36 | 180 |
| 11 | 9 | 99 | -3 | 9 | 81 |
| 17 | 5 | 85 | 3 | 9 | 45 |
| 20 | 4 | 80 | 6 | 36 | 144 |
| 24 | 3 | 72 | 10 | 100 | 300 |
| 32 | 1 | 32 | 18 | 324 | 324 |
| | 30 | 420 | | | 1374 |

$$N = 30, \sum_{i=1}^7 f_i x_i = 420, \sum_{i=1}^7 f_i (x_i - \bar{x})^2 = 1374$$

Therefore

$$\bar{x} = \frac{\sum_{i=1}^7 f_i x_i}{N} = \frac{1}{30} \times 420 = 14$$

Hence

$$\begin{aligned} \text{variance } (\sigma^2) &= \frac{1}{N} \sum_{i=1}^7 f_i (x_i - \bar{x})^2 \\ &= \frac{1}{30} \times 1374 = 45.8 \end{aligned}$$

and

$$\text{Standard deviation } (\sigma) = \sqrt{45.8} = 6.77$$

13.5.3 Standard deviation of a continuous frequency distribution The given continuous frequency distribution can be represented as a discrete frequency distribution by replacing each class by its mid-point. Then, the standard deviation is calculated by the technique adopted in the case of a discrete frequency distribution.

If there is a frequency distribution of n classes each class defined by its mid-point x_i with frequency f_i , the standard deviation will be obtained by the formula

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

where \bar{x} is the mean of the distribution and $N = \sum_{i=1}^n f_i$.

Another formula for standard deviation We know that

$$\begin{aligned} \text{Variance } (\sigma^2) &= \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_{i=1}^n f_i (x_i^2 + \bar{x}^2 - 2\bar{x} x_i) \\ &= \frac{1}{N} \left[\sum_{i=1}^n f_i x_i^2 + \sum_{i=1}^n \bar{x}^2 f_i - \sum_{i=1}^n 2\bar{x} f_i x_i \right] \\ &= \frac{1}{N} \left[\sum_{i=1}^n f_i x_i^2 + \bar{x}^2 \sum_{i=1}^n f_i - 2\bar{x} \sum_{i=1}^n x_i f_i \right] \end{aligned}$$

$$\begin{aligned} &= \frac{1}{N} \sum_{i=1}^n f_i x_i^2 + \bar{x}^2 N - 2\bar{x} \cdot N \bar{x} \quad \left[\text{Here } \frac{1}{N} \sum_{i=1}^n x_i f_i = \bar{x} \text{ or } \sum_{i=1}^n x_i f_i = N\bar{x} \right] \\ &= \frac{1}{N} \sum_{i=1}^n f_i x_i^2 + \bar{x}^2 N - 2\bar{x}^2 N = \frac{1}{N} \sum_{i=1}^n f_i x_i^2 - \bar{x}^2 \\ \text{or} \quad \sigma^2 &= \frac{1}{N} \sum_{i=1}^n f_i x_i^2 - \left(\frac{\sum_{i=1}^n f_i x_i}{N} \right)^2 = \frac{1}{N^2} \left[N \sum_{i=1}^n f_i x_i^2 - \left(\sum_{i=1}^n f_i x_i \right)^2 \right] \end{aligned}$$

Thus, standard deviation (σ) = $\frac{1}{N} \sqrt{N \sum_{i=1}^n f_i x_i^2 - \left(\sum_{i=1}^n f_i x_i \right)^2}$... (3)

Example 10 Calculate the mean, variance and standard deviation for the following distribution :

| | | | | | | | |
|-----------|-------|-------|-------|-------|-------|-------|--------|
| Class | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 |
| Frequency | 3 | 7 | 12 | 15 | 8 | 3 | 2 |

Solution From the given data, we construct the following Table 13.9.

Table 13.9

| Class | Frequency (f_i) | Mid-point (x_i) | $f_i x_i$ | $(x_i - \bar{x})^2$ | $f_i (x_i - \bar{x})^2$ |
|--------|------------------------|------------------------|-----------|---------------------|-------------------------|
| 30-40 | 3 | 35 | 105 | 729 | 2187 |
| 40-50 | 7 | 45 | 315 | 289 | 2023 |
| 50-60 | 12 | 55 | 660 | 49 | 588 |
| 60-70 | 15 | 65 | 975 | 9 | 135 |
| 70-80 | 8 | 75 | 600 | 169 | 1352 |
| 80-90 | 3 | 85 | 255 | 529 | 1587 |
| 90-100 | 2 | 95 | 190 | 1089 | 2178 |
| | 50 | | 3100 | | 10050 |

Thus Mean $\bar{x} = \frac{1}{N} \sum_{i=1}^7 f_i x_i = \frac{3100}{50} = 62$

$$\begin{aligned} \text{Variance } (\sigma^2) &= \frac{1}{N} \sum_{i=1}^7 f_i (x_i - \bar{x})^2 \\ &= \frac{1}{50} \times 10050 = 201 \end{aligned}$$

and Standard deviation $(\sigma) = \sqrt{201} = 14.18$

Example 11 Find the standard deviation for the following data :

| | | | | | |
|-------|---|----|----|----|----|
| x_i | 3 | 8 | 13 | 18 | 23 |
| f_i | 7 | 10 | 15 | 10 | 6 |

Solution Let us form the following Table 13.10:

Table 13.10

| x_i | f_i | $f_i x_i$ | x_i^2 | $f_i x_i^2$ |
|-------|-------|-----------|---------|-------------|
| 3 | 7 | 21 | 9 | 63 |
| 8 | 10 | 80 | 64 | 640 |
| 13 | 15 | 195 | 169 | 2535 |
| 18 | 10 | 180 | 324 | 3240 |
| 23 | 6 | 138 | 529 | 3174 |
| | 48 | 614 | | 9652 |

Now, by formula (3), we have

$$\begin{aligned} \sigma &= \frac{1}{N} \sqrt{N \sum f_i x_i^2 - (\sum f_i x_i)^2} \\ &= \frac{1}{48} \sqrt{48 \times 9652 - (614)^2} \\ &= \frac{1}{48} \sqrt{463296 - 376996} \end{aligned}$$

$$= \frac{1}{48} \times 293.77 = 6.12$$

Therefore, Standard deviation (σ) = 6.12

13.5.4. Shortcut method to find variance and standard deviation Sometimes the values of x_i in a discrete distribution or the mid points x_i of different classes in a continuous distribution are large and so the calculation of mean and variance becomes tedious and time consuming. By using step-deviation method, it is possible to simplify the procedure.

Let the assumed mean be 'A' and the scale be reduced to $\frac{1}{h}$ times (h being the width of class-intervals). Let the step-deviations or the new values be y_i .

i.e. $y_i = \frac{x_i - A}{h}$ or $x_i = A + hy_i$... (1)

We know that $\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{N}$... (2)

Replacing x_i from (1) in (2), we get

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^n f_i (A + hy_i)}{N} \\ &= \frac{1}{N} \left(\sum_{i=1}^n f_i A + \sum_{i=1}^n h f_i y_i \right) = \frac{1}{N} \left(A \sum_{i=1}^n f_i + h \sum_{i=1}^n f_i y_i \right) \\ &= A \cdot \frac{N}{N} + h \frac{\sum_{i=1}^n f_i y_i}{N} \quad \left(\text{because } \sum_{i=1}^n f_i = N \right) \end{aligned}$$

Thus $\bar{x} = A + h \bar{y}$... (3)

Now Variance of the variable x , $\sigma_x^2 = \frac{1}{N} \sum_{i=1}^n f_i (x_i - \bar{x})^2$

$$= \frac{1}{N} \sum_{i=1}^n f_i (A + hy_i - A - h \bar{y})^2 \quad (\text{Using (1) and (3)})$$

$$\begin{aligned}
 &= \frac{1}{N} \sum_{i=1}^n f_i h^2 (y_i - \bar{y})^2 \\
 &= \frac{h^2}{N} \sum_{i=1}^n f_i (y_i - \bar{y})^2 = h^2 \times \text{variance of the variable } y_i
 \end{aligned}$$

i.e. $\sigma_x^2 = h^2 \sigma_y^2$

or $\sigma_x = h \sigma_y$... (4)

From (3) and (4), we have

$$\sigma_x = \frac{h}{N} \sqrt{N \sum_{i=1}^n f_i y_i^2 - \left(\sum_{i=1}^n f_i y_i \right)^2} \quad \dots (5)$$

Let us solve Example 11 by the short-cut method and using formula (5)

Examples 12 Calculate mean, variance and standard deviation for the following distribution.

| | | | | | | | |
|-----------|-------|-------|-------|-------|-------|-------|--------|
| Classes | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 |
| Frequency | 3 | 7 | 12 | 15 | 8 | 3 | 2 |

Solution Let the assumed mean $A = 65$. Here $h = 10$
 We obtain the following Table 13.11 from the given data :

Table 13.11

| Class | Frequency | Mid-point | $y_i = \frac{x_i - 65}{10}$ | y_i^2 | $f_i y_i$ | $f_i y_i^2$ |
|--------|-----------|-----------|-----------------------------|---------|-----------|-------------|
| | f_i | x_i | | | | |
| 30-40 | 3 | 35 | -3 | 9 | -9 | 27 |
| 40-50 | 7 | 45 | -2 | 4 | -14 | 28 |
| 50-60 | 12 | 55 | -1 | 1 | -12 | 12 |
| 60-70 | 15 | 65 | 0 | 0 | 0 | 0 |
| 70-80 | 8 | 75 | 1 | 1 | 8 | 8 |
| 80-90 | 3 | 85 | 2 | 4 | 6 | 12 |
| 90-100 | 2 | 95 | 3 | 9 | 6 | 18 |
| | N=50 | | | | -15 | 105 |

Therefore $\bar{x} = A + \frac{\sum f_i y_i}{50} \times h = 65 - \frac{15}{50} \times 10 = 62$

Variance
$$\begin{aligned}\sigma^2 &= \frac{h^2}{N^2} \left[N \sum f_i y_i^2 - \left(\sum f_i y_i \right)^2 \right] \\ &= \frac{(10)^2}{(50)^2} \left[50 \times 105 - (-15)^2 \right] \\ &= \frac{1}{25} [5250 - 225] = 201\end{aligned}$$

and standard deviation $(\sigma) = \sqrt{201} = 14.18$

EXERCISE 13.2

Find the mean and variance for each of the data in Exercises 1 to 5.

- 6, 7, 10, 12, 13, 4, 8, 12
- First n natural numbers
- First 10 multiples of 3

4.

| | | | | | | | |
|-------|---|----|----|----|----|----|----|
| x_i | 6 | 10 | 14 | 18 | 24 | 28 | 30 |
| f_i | 2 | 4 | 7 | 12 | 8 | 4 | 3 |

5.

| | | | | | | | |
|-------|----|----|----|----|-----|-----|-----|
| x_i | 92 | 93 | 97 | 98 | 102 | 104 | 109 |
| f_i | 3 | 2 | 3 | 2 | 6 | 3 | 3 |

6. Find the mean and standard deviation using short-cut method.

| | | | | | | | | | |
|-------|----|----|----|----|----|----|----|----|----|
| x_i | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 |
| f_i | 2 | 1 | 12 | 29 | 25 | 12 | 10 | 4 | 5 |

Find the mean and variance for the following frequency distributions in Exercises 7 and 8.

7.

| | | | | | | | |
|-------------|------|-------|-------|--------|---------|---------|---------|
| Classes | 0-30 | 30-60 | 60-90 | 90-120 | 120-150 | 150-180 | 180-210 |
| Frequencies | 2 | 3 | 5 | 10 | 3 | 5 | 2 |

8.

| | | | | | |
|-------------|------|-------|-------|-------|-------|
| Classes | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 |
| Frequencies | 5 | 8 | 15 | 16 | 6 |

9. Find the mean, variance and standard deviation using short-cut method

| | | | | | | | | | |
|-----------------|-------|-------|-------|-------|-------|--------|---------|---------|---------|
| Height in cms | 70-75 | 75-80 | 80-85 | 85-90 | 90-95 | 95-100 | 100-105 | 105-110 | 110-115 |
| No. of children | 3 | 4 | 7 | 7 | 15 | 9 | 6 | 6 | 3 |

10. The diameters of circles (in mm) drawn in a design are given below:

| | | | | | |
|----------------|-------|-------|-------|-------|-------|
| Diameters | 33-36 | 37-40 | 41-44 | 45-48 | 49-52 |
| No. of circles | 15 | 17 | 21 | 22 | 25 |

Calculate the standard deviation and mean diameter of the circles.

[**Hint** First make the data continuous by making the classes as 32.5-36.5, 36.5-40.5, 40.5-44.5, 44.5 - 48.5, 48.5 - 52.5 and then proceed.]

Miscellaneous Examples

Example 13 The variance of 20 observations is 5. If each observation is multiplied by 2, find the new variance of the resulting observations.

Solution Let the observations be x_1, x_2, \dots, x_{20} and \bar{x} be their mean. Given that variance = 5 and $n = 20$. We know that

$$\text{Variance } (\sigma^2) = \frac{1}{n} \sum_{i=1}^{20} (x_i - \bar{x})^2, \text{ i.e., } 5 = \frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})^2$$

or

$$\sum_{i=1}^{20} (x_i - \bar{x})^2 = 100$$

If each observation is multiplied by 2, and the new resulting observations are y_i , then

$$y_i = 2x_i \text{ i.e., } x_i = \frac{1}{2} y_i \quad \dots (1)$$


Therefore
$$\bar{y} = \frac{1}{n} \sum_{i=1}^{20} y_i = \frac{1}{20} \sum_{i=1}^{20} 2x_i = 2 \cdot \frac{1}{20} \sum_{i=1}^{20} x_i$$

i.e.
$$\bar{y} = 2\bar{x} \quad \text{or} \quad \bar{x} = \frac{1}{2}\bar{y}$$

Substituting the values of x_i and \bar{x} in (1), we get

$$\sum_{i=1}^{20} \left(\frac{1}{2} y_i - \frac{1}{2} \bar{y} \right)^2 = 100, \text{ i.e., } \sum_{i=1}^{20} (y_i - \bar{y})^2 = 400$$

Thus the variance of new observations $= \frac{1}{20} \times 400 = 20 = 2^2 \times 5$

 **Note** The reader may note that if each observation is multiplied by a constant k , the variance of the resulting observations becomes k^2 times the original variance.

Example14 The mean of 5 observations is 4.4 and their variance is 8.24. If three of the observations are 1, 2 and 6, find the other two observations.

Solution Let the other two observations be x and y .

Therefore, the series is 1, 2, 6, x , y .

Now
$$\text{Mean } \bar{x} = 4.4 = \frac{1+2+6+x+y}{5}$$

or
$$22 = 9 + x + y$$

Therefore
$$x + y = 13 \quad \dots (1)$$

Also
$$\text{variance} = 8.24 = \frac{1}{n} \sum_{i=1}^5 (x_i - \bar{x})^2$$

i.e.
$$8.24 = \frac{1}{5} \left[(3.4)^2 + (2.4)^2 + (1.6)^2 + x^2 + y^2 - 2 \times 4.4(x+y) + 2 \times (4.4)^2 \right]$$

or
$$41.20 = 11.56 + 5.76 + 2.56 + x^2 + y^2 - 8.8 \times 13 + 38.72$$

Therefore
$$x^2 + y^2 = 97 \quad \dots (2)$$

But from (1), we have

$$x^2 + y^2 + 2xy = 169 \quad \dots (3)$$

From (2) and (3), we have

$$2xy = 72 \quad \dots (4)$$

Subtracting (4) from (2), we get

$$x^2 + y^2 - 2xy = 97 - 72 \text{ i.e. } (x - y)^2 = 25$$

$$\text{or } x - y = \pm 5 \quad \dots (5)$$

So, from (1) and (5), we get

$$x = 9, y = 4 \text{ when } x - y = 5$$

$$\text{or } x = 4, y = 9 \text{ when } x - y = -5$$

Thus, the remaining observations are 4 and 9.

Example 15 If each of the observation x_1, x_2, \dots, x_n is increased by 'a', where a is a negative or positive number, show that the variance remains unchanged.

Solution Let \bar{x} be the mean of x_1, x_2, \dots, x_n . Then the variance is given by

$$\sigma_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

If 'a' is added to each observation, the new observations will be

$$y_i = x_i + a \quad \dots (1)$$

Let the mean of the new observations be \bar{y} . Then


$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n (x_i + a) \\ &= \frac{1}{n} \left[\sum_{i=1}^n x_i + \sum_{i=1}^n a \right] = \frac{1}{n} \sum_{i=1}^n x_i + \frac{na}{n} = \bar{x} + a \end{aligned}$$

$$\text{i.e. } \bar{y} = \bar{x} + a \quad \dots (2)$$

Thus, the variance of the new observations

$$\begin{aligned} \sigma_2^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (x_i + a - \bar{x} - a)^2 \quad [\text{Using (1) and (2)}] \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \sigma_1^2 \end{aligned}$$

Thus, the variance of the new observations is same as that of the original observations.

 **Note** We may note that adding (or subtracting) a positive number to (or from) each observation of a group does not affect the variance.

Example 16 The mean and standard deviation of 100 observations were calculated as 40 and 5.1, respectively by a student who took by mistake 50 instead of 40 for one observation. What are the correct mean and standard deviation?

Solution Given that number of observations (n) = 100

Incorrect mean (\bar{x}) = 40,

Incorrect standard deviation (σ) = 5.1

We know that $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

$$\text{i.e.} \quad 40 = \frac{1}{100} \sum_{i=1}^{100} x_i \quad \text{or} \quad \sum_{i=1}^{100} x_i = 4000$$

i.e. Incorrect sum of observations = 4000

Thus the correct sum of observations = Incorrect sum – 50 + 40
 $= 4000 - 50 + 40 = 3990$

$$\text{Hence} \quad \text{Correct mean} = \frac{\text{correct sum}}{100} = \frac{3990}{100} = 39.9$$

$$\begin{aligned} \text{Also} \quad \text{Standard deviation } \sigma &= \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \frac{1}{n^2} \left(\sum_{i=1}^n x_i \right)^2} \\ &= \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2} \end{aligned}$$

$$\text{i.e.} \quad 5.1 = \sqrt{\frac{1}{100} \times \text{Incorrect} \sum_{i=1}^n x_i^2 - (40)^2}$$

$$\text{or} \quad 26.01 = \frac{1}{100} \times \text{Incorrect} \sum_{i=1}^n x_i^2 - 1600$$

$$\text{Therefore} \quad \text{Incorrect} \sum_{i=1}^n x_i^2 = 100 (26.01 + 1600) = 162601$$

$$\begin{aligned} \text{Now} \quad \text{Correct} \sum_{i=1}^n x_i^2 &= \text{Incorrect} \sum_{i=1}^n x_i^2 - (50)^2 + (40)^2 \\ &= 162601 - 2500 + 1600 = 161701 \end{aligned}$$

Therefore Correct standard deviation

$$\begin{aligned}
 &= \sqrt{\frac{\text{Correct } \sum x_i^2}{n} - (\text{Correct mean})^2} \\
 &= \sqrt{\frac{161701}{100} - (39.9)^2} \\
 &= \sqrt{1617.01 - 1592.01} = \sqrt{25} = 5
 \end{aligned}$$

Miscellaneous Exercise On Chapter 13

1. The mean and variance of eight observations are 9 and 9.25, respectively. If six of the observations are 6, 7, 10, 12, 12 and 13, find the remaining two observations.
2. The mean and variance of 7 observations are 8 and 16, respectively. If five of the observations are 2, 4, 10, 12, 14. Find the remaining two observations.
3. The mean and standard deviation of six observations are 8 and 4, respectively. If each observation is multiplied by 3, find the new mean and new standard deviation of the resulting observations.
4. Given that \bar{x} is the mean and σ^2 is the variance of n observations x_1, x_2, \dots, x_n . Prove that the mean and variance of the observations $ax_1, ax_2, ax_3, \dots, ax_n$ are $a\bar{x}$ and $a^2\sigma^2$, respectively, ($a \neq 0$).
5. The mean and standard deviation of 20 observations are found to be 10 and 2, respectively. On rechecking, it was found that an observation 8 was incorrect. Calculate the correct mean and standard deviation in each of the following cases:
(i) If wrong item is omitted. (ii) If it is replaced by 12.
6. The mean and standard deviation of a group of 100 observations were found to be 20 and 3, respectively. Later on it was found that three observations were incorrect, which were recorded as 21, 21 and 18. Find the mean and standard deviation if the incorrect observations are omitted.

Summary

- ◆ **Measures of dispersion** Range, Quartile deviation, mean deviation, variance, standard deviation are measures of dispersion.

Range = Maximum Value – Minimum Value

- ◆ **Mean deviation for ungrouped data**

$$\text{M.D. } (\bar{x}) = \frac{\sum |x_i - \bar{x}|}{n}, \quad \text{M.D. } (M) = \frac{\sum |x_i - M|}{n}$$

◆ **Mean deviation for grouped data**

$$\text{M.D.}(\bar{x}) = \frac{\sum f_i |x_i - \bar{x}|}{N}, \quad \text{M.D.}(M) = \frac{\sum f_i |x_i - M|}{N}, \text{ where } N = \sum f_i$$

◆ **Variance and standard deviation for ungrouped data**

$$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2, \quad \sigma = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

◆ **Variance and standard deviation of a discrete frequency distribution**

$$\sigma^2 = \frac{1}{N} \sum f_i (x_i - \bar{x})^2, \quad \sigma = \sqrt{\frac{1}{N} \sum f_i (x_i - \bar{x})^2}$$

◆ **Variance and standard deviation of a continuous frequency distribution**

$$\sigma^2 = \frac{1}{N} \sum f_i (x_i - \bar{x})^2, \quad \sigma = \frac{1}{N} \sqrt{N \sum f_i x_i^2 - (\sum f_i x_i)^2}$$

◆ **Shortcut method to find variance and standard deviation.**

$$\sigma^2 = \frac{h^2}{N^2} \left[N \sum f_i y_i^2 - (\sum f_i y_i)^2 \right], \quad \sigma = \frac{h}{N} \sqrt{N \sum f_i y_i^2 - (\sum f_i y_i)^2},$$

$$\text{where } y_i = \frac{x_i - A}{h}$$

Historical Note

‘Statistics’ is derived from the Latin word ‘status’ which means a political state. This suggests that statistics is as old as human civilisation. In the year 3050 B.C., perhaps the first census was held in Egypt. In India also, about 2000 years ago, we had an efficient system of collecting administrative statistics, particularly, during the regime of Chandra Gupta Maurya (324-300 B.C.). The system of collecting data related to births and deaths is mentioned in Kautilya’s *Arthshastra* (around 300 B.C.) A detailed account of administrative surveys conducted during Akbar’s regime is given in *Ain-I-Akbari* written by Abul Fazl.

Captain John Graunt of London (1620-1674) is known as father of vital statistics due to his studies on statistics of births and deaths. Jacob Bernoulli (1654-1705) stated the Law of Large numbers in his book “*Ars Conjectandi*”, published in 1713.

The theoretical development of statistics came during the mid seventeenth century and continued after that with the introduction of theory of games and chance (i.e., probability). Francis Galton (1822-1921), an Englishman, pioneered the use of statistical methods, in the field of Biometry. Karl Pearson (1857-1936) contributed a lot to the development of statistical studies with his discovery of *Chi square test* and foundation of *statistical laboratory* in England (1911). Sir Ronald A. Fisher (1890-1962), known as the Father of modern statistics, applied it to various diversified fields such as Genetics, Biometry, Education, Agriculture, etc.



© NCERT
not to be republished