# 13 Descriptive Statistics

In previous classes we have learnt about statistics as classification of data, graphical representation of data, mean, median and mode of ungrouped and grouped data.

Statistics is a branch of mathematics which deals with collection, organization, analysis, interpretation and presentation of data.

The word statistics seems reams to have derived from the Latin word status which means a political state. Originally statistics was simply the collection of numerical data on some aspects of life of the people useful to the government. However, with the passage of time its scope broadend. Today statistics means collection of facts or information concerning almost every aspect of life such as science, medicine, economics, psychology, politics and marketing.

The study of statistics can be categorized into two types:

**1.** Descriptive statistics  **2.** Inferential statistics

In this chapter we shall learn about descriptive statistics only.

## DESCRIPTIVE STATISTICS

Descriptive statistics are used to describe or summarize data in useful and meaningful ways. Descriptive statistics can be useful for two purposes.

(*i*) To provide basic information about variables in a data set.

(*ii*) To highlight potential relationship between variables.

There are four major types of descriptive statistics:

1. Measures of central tendency: Mean, Median and Mode
2. Measures of dispersion: Deviation, Variance, Standard deviation, Skewness, Kurtosis
3. Measures of position: Percentile rank, Quartile rank
4. Measures of association : Correlation

## MEASURES OF DISPERSION

### Describing the Dispersion

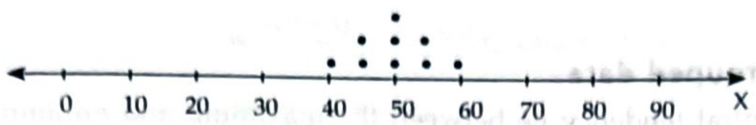
You have already studied the locational statistics—which give us a sort of central value around which the values of the variable are located. These measures of central tendency give a rough idea of where data points are centred. However, in order to interpret the data better, we should also know *how far these values spread around the central value.*

Consider an example of a cricket team and its batting performance in last 9 one day matches. Let us assume that batsman A scored 60, 55, 50, 50, 40, 45, 55, 45, 50; batsman B scored 70, 30, 60, 20, 50, 90, 40, 80, 10; batsman C scored 20, 25, 15, 25, 15, 20, 20 in seven innings and did not get to bat in 2 innings. The mean and median are :

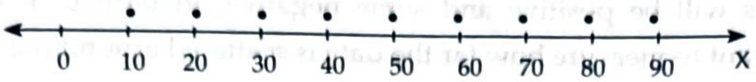
| | *A* | *B* | *C* |
|---|---|---|---|
| *Mean* | 50 | 50 | 20 |
| *Median* | 50 | 50 | 20 |

This tells us that average performance of batsman A and B is the same, and much better than that of C. But this is only part of the story. Now let us plot these scores as dots on a number line.

For batsman A :




For batsman B :




For batsman C :


These diagrams show that average score of A and B is the same, the score distribution of batsman A is more 'compact' than that of B whose score distribution is comparatively *dispersed* (*scattered*) widely. When A goes out to bat, you are quite sure that he will score around a half century, but when batsman B goes out to bat, you keep your fingers crossed. You don't know whether he will be out for a duck or whether he will hit a century. Though his average is also half a century. As far as batsman C is concerned, though his average is low, he is quite *reliable* when he goes out to bat, you are pretty certain that he will score "around" 20 runs.

This leads us to second set of descriptives—**measures of variability** or **spread of distribution**.

**Definition.** *Dispersion indicates the extent to which the individual measures differ from an average.*

## DIFFERENT METHODS OF MEASURING DISPERSION

There are many ways in which the dispersion *i.e.* the spread of the data can be measured. These include :

(*i*) Range (*ii*) Quartile deviation

(*iii*) Mean deviation (*iv*) Standard deviation

Here, we shall study the mean deviation about the mean, mean deviation about the median, and standard deviation.

### Range

The difference between the largest value and smallest value of a data series is called its **range**. Thus the range indicates the total spread of the data *i.e.* 100% values lie within the range.

In the example given in the article Describing the Dispersion,

for batsman A, range = maximum value – minimum value

= 60 – 40 = 20.

For batsman B, range = 90 – 10 = 80

For batsman C, range = 25 – 15 = 10.

For grouped data, range is defined as the difference between the upper boundary of the highest class and lower boundary of the lowest class.

### Quartile Deviation (Q.D.)

Quartile deviation is a statistic that measures the deviation in the middle of the data. Quartile deviation is also referred as the semi-interquartile range ($Q_3 - Q_1$). The formula for quartile deviation of the data is

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}.$$

We have learnt to calculate quartiles for grouped and ungrouped data in the previous part of this chapter.

## MEAN DEVIATION

### Mean deviation for ungrouped data

Recall that measures of central tendency lie between the maximum and minimum values of a set of observations. Thus, if we consider the deviations $x_i - a$ from such a measure $a$, then some of the values of deviations will be positive and some negative. In particular, we know that $\Sigma (x_i - \bar{x}) = 0$. Thus, if we want to measure how far the data is scattered around the mean, we may take absolute values of deviations $x_i - \bar{x}$ and then find their average.

Thus, mean deviation about the mean ($\bar{x}$) is

$$\text{M.D.}(\bar{x}) = \frac{\Sigma |x_i - \bar{x}|}{n}, \text{ where } n \text{ is number of observations.}$$

Similarly, mean deviation about the median (M) is

$$\text{M.D.}(\text{M}) = \frac{\Sigma |x_i - \text{M}|}{n}, \text{ where } n \text{ is number of observations.}$$

Thus, in the example taken above, for batsman A the deviation around the mean (50) for 9 values are 10, 5, 0, 0, −10, −5, 5, −5, 0. The absolute values of these deviations are 10, 5, 0, 0, 10, 5, 5, 5, 0, whose sum is 40.

Hence, mean deviation about the mean $= \frac{40}{9} = 4.4$ (approx.)

Similarly, for batsman B, the deviations around the mean (50) are 20, −20, 10, −30, 0, 40, −10, 30, −40; absolute deviations are 20, 20, 10, 30, 0, 40, 10, 30, 40; their sum is 200. Hence, mean deviation about the mean $= \frac{200}{9} = 22.2$ (approx). We immediately notice that dispersion value for batsman B is much higher than that of A.

## ILLUSTRATIVE EXAMPLES

**Example 1.** *Find out the range of salaries for staff of a school:*

| *Staff* | *Principal* | *Senior teacher* | *Junior teacher* | *Other workers* |
|---|---|---|---|---|
| *Salary (in ₹)* | *100000* | *70000* | *45000* | *25000* |

**Solution.** Range = Maximum value − Minimum value

= ₹100000 − ₹25000 = ₹75000

**Example 2.** *The marks of 9 students in a test were 13, 17, 20, 5, 3, 3, 18, 15 and 20. Calculate the quartile deviation.*

**Solution.** Arranging the data in ascending order, we get

3, 3, 5, 13, 15, 17, 18, 20, 20

Here, $n = 9$

Lower or first quartile $Q_1 = \frac{n+1}{4}$ th observation

$= \frac{9+1}{2}$ th *i.e.* 2.5 th observation

$$\therefore \quad Q_1 = \frac{\text{2nd observation} + \text{3rd observation}}{2} = \frac{3+5}{2} = 4$$

Upper or third quartile $Q_3 = \frac{3(n+1)}{4}$ th observation

$$= \frac{3(9+1)}{4} \text{ th } i.e. \text{ 7.5 th observation}$$

$$\therefore \quad Q_3 = \frac{\text{7th observation + 8th observation}}{2}$$

$$= \frac{18+20}{2} = 19$$

Here, $Q_1 = 4$, $Q_3 = 19$

$$\therefore \quad \text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{19-4}{2} = 7.5$$

✗ **Example 3.** *The following table gives information regarding weekly income of labourers working at a dam site:*

| *Income (in ₹)* | 600 – 700 | 700 – 800 | 800 – 900 | 900 – 1000 | 1000 – 1100 | 1100 – 1200 | 1200 – 1300 |
|---|---|---|---|---|---|---|---|
| *No. of labourers* | 40 | 68 | 86 | 120 | 90 | 40 | 26 |

*Calculate the quartile deviation.*

**Solution.** Here, the classes are already continuous and of uniform width. The cumulative frequency table is as under:

| *Weekly income (in ₹)* | *No. of labourers* | *Cumulative frequency* |
|---|---|---|
| 600 – 700 | 40 | 40 |
| 700 – 800 | 68 | 108 |
| **800 – 900** | **86** | **194** |
| 900 – 1000 | 120 | 314 |
| 1000 – 1100 | 90 | 404 |
| 1100 – 1200 | 40 | 444 |
| 1200 – 1300 | 26 | 470 |

Here, $n = 470$

Lower quartile ($Q_1$) is the $\frac{n}{4}$ th *i.e.* $\frac{470}{4}$ th *i.e.* 117.5th value, which lies in class 800 – 900

$$\therefore \quad Q_1 = l + \frac{\frac{n}{4} - C}{f} \times h = 800 + \frac{117.5 - 108}{86} \times 100$$

$$= 800 + 11.01$$

$$= 811.01 \; i.e. \text{ ₹ } 811$$

Upper quartile ($Q_3$) is the $\frac{3n}{4}$ th *i.e.* 352.5th value, which lies in class 1000 – 1100

$$\therefore \quad Q_3 = l + \frac{\frac{3n}{4} - C}{f} \times h = 1000 + \frac{352.5 - 314}{90} \times 100$$

$$= 1000 + 42.78$$

$$= 1042.78 \; i.e. \text{ ₹ } 1042.78$$

Here, $Q_1$ = ₹811 and $Q_3$ = ₹1042.78

$$\therefore \quad \text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{1042.78 - 811}{2} = ₹115.89$$

Hence, the quartile deviation of the given data is ₹115.89.

**Example 4.** *Find the mean deviation about the mean for the following series :*

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

**Solution.** The data values in the ascending order are

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

Here, the total number of observations = 20, which is even.

$$\therefore \quad \text{Mean} = \frac{0+1+2+3+3+3+4+5+8+9+11+12+15+16+17+17+17+18+19+20}{20} = \frac{200}{20} = 10.$$

The absolute values of the deviations from the mean *i.e.* $|x_i - 10|$ are

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

Their sum = 124.

$$\therefore \quad \text{Mean deviation about the mean} = \frac{\Sigma|x_i - 10|}{n} = \frac{124}{20} = 6.2$$

## Mean deviation for grouped data

Data can be grouped in two ways :

(*i*) Discrete frequency distribution

(*ii*) Continuous frequency distribution

Let us discuss the mean deviation in these distributions one by one.

## Discrete frequency distribution

Let the given data consist of $m$ distinct values $x_1, x_2, \ldots, x_m$ with frequencies $f_1, f_2, \ldots, f_m$ respectively. This data can be represented in tabular form as given below and is called discrete frequency distribution.

| $x$ | $x_1$ | $x_2$ | ... | $x_m$ |
|---|---|---|---|---|
| $f$ | $f_1$ | $f_2$ | ... | $f_m$ |

## Mean deviation about the mean/median

To find the mean deviation about the mean ($\bar{x}$), first we find the absolute values of deviations $x_i - \bar{x}$, and then use the formula

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

After finding median (M), we obtain the mean deviation from the median by using the formula

$$\text{M.D.}(M) = \frac{\Sigma f_i|x_i - M|}{\Sigma f_i} = \frac{1}{n}\Sigma f_i|x_i - M|.$$

**Example 5.** *Find the mean deviation about the mean for the following data :*

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

**Solution.** To calculate mean, we require $f_i x_i$ values; then to find mean deviation, we will require $|x_i - \bar{x}|$ values and $f_i|x_i - \bar{x}|$ values. Hence, we make the following table. (Note that 4th column is added after calculating $\bar{x}$, then 5th column is added).

| $x_i$ | $f_i$ | $f_ix_i$ | $\|x_i - \bar{x}\|$ | $f_i\|x_i - \bar{x}\|$ |
|---|---|---|---|---|
| 10 | 4 | 40 | 40 | 160 |
| 30 | 24 | 720 | 20 | 480 |
| 50 | 28 | 1400 | 0 | 0 |
| 70 | 16 | 1120 | 20 | 320 |
| 90 | 8 | 720 | 40 | 320 |
| | 80 | 4000 | | 1280 |

Here $n = \Sigma f_i = 80$ and $\Sigma f_i x_i = 4000$

$$\therefore \quad \bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{400}{80} = 50.$$

Mean deviation about the mean,

$$\text{M.D.}(\bar{x}) = \frac{\Sigma f_i \,|x_i - \bar{x}|}{n} = \frac{1280}{80} = 16.$$

**Example 6.** *Find the mean deviation about the median for the following data :*

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

**Solution.** The given values are already in ascending order. We construct the following table to calculate the median.

| $x_i$ | 5 | 7 | 9 | 10 | 12 | 15 |
|---|---|---|---|---|---|---|
| $f_i$ | 8 | 6 | 2 | 2 | 2 | 6 |
| *c.f.* | 8 | 14 | 16 | 18 | 20 | 26 |

Here $n = 26$, which is even. So median is the average of 13th and 14th observations

$$\therefore \quad \text{Median M} = \frac{\text{13th observation + 14th observation}}{2} = \frac{7+7}{2} = 7.$$

Now we calculate the M.D. (M) by constructing the following table:

| $\|x_i - \text{M}\|$ | 2 | 0 | 2 | 3 | 5 | 8 |
|---|---|---|---|---|---|---|
| $f_i$ | 8 | 6 | 2 | 2 | 2 | 6 |
| $f_i\|x_i - \text{M}\|$ | 16 | 0 | 4 | 6 | 10 | 48 |

Here $n = \Sigma f_i = 26$, $\Sigma f_i\,|x_i - \text{M}| = 84$

$$\therefore \quad \text{M.D. (M)} = \frac{1}{n}\,\Sigma f_i\,|x_i - \text{M}| = \frac{84}{26} = \frac{42}{13} = 3.23 \text{ (approx.)}$$

## Continuous frequency distribution

In a continuous frequency distribution, the data is classified into different class intervals without gaps along with their respective frequencies. To calculate the mean deviation, first we calculate the mean or median as the case may be. Then to calculate deviations, we assume that the frequency in each class is centred at its mid-point. We take this mid-point's deviation from mean or median, and proceed as in discrete frequency distribution to calculate M.D. ($\bar{x}$) or M.D. (M). This will be clear from the following illustrative examples.

**Note :** If the classes are discontinuous, then first convert them into continuous classes.

**Example 7.** *Calculate the mean deviation about the mean for the following data :*

| *Income per day* | 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 |

**Solution.** We construct the following table. (5th and 6th columns are filled after calculating the mean.)

| *Income per day* | *Number of persons* $f_i$ | *Mid-points* $x_i$ | $f_i x_i$ | $\lvert x_i - \bar{x} \rvert$ | $f_i \lvert x_i - \bar{x} \rvert$ |
|---|---|---|---|---|---|
| 0–100 | 4 | 50 | 200 | 308 | 1232 |
| 100–200 | 8 | 150 | 1200 | 208 | 1664 |
| 200–300 | 9 | 250 | 2250 | 108 | 972 |
| 300–400 | 10 | 350 | 3500 | 8 | 80 |
| 400–500 | 7 | 450 | 3150 | 92 | 644 |
| 500–600 | 5 | 550 | 2750 | 192 | 960 |
| 600–700 | 4 | 650 | 2600 | 292 | 1168 |
| 700–800 | 3 | 750 | 2250 | 392 | 1176 |
| Total | 50 | | 17900 | | 7896 |

Here $n = \Sigma f_i = 50, \Sigma f_i x_i = 17900$

$$\therefore \quad \text{Mean} = \bar{x} = \frac{1}{n} \Sigma f_i x_i = \frac{17900}{50} = 358$$

$$\text{M.D.} (\bar{x}) = \frac{1}{n} \Sigma f_i \lvert x_i - \bar{x} \rvert = \frac{7896}{50} = 157.92$$

**Example 8.** *Find the mean deviation about the mean for the following data:*

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

**Solution.** Taking assumed mean $a = 120$ and class size $= c = 10$, we construct the following table. Note that 6th and 7th columns are filled in after calculating the mean of the given data.

| *Height in cm* | *No. of boys* $f_i$ | *Mid-points* $x_i$ | $u_i = \frac{x_i - 120}{10}$ | $f_i u_i$ | $\lvert x_i - \bar{x} \rvert$ | $f_i \lvert x_i - \bar{x} \rvert$ |
|---|---|---|---|---|---|---|
| 95–105 | 9 | 100 | –2 | –18 | 25.3 | 227.7 |
| 105–115 | 13 | 110 | –1 | –13 | 15.3 | 198.9 |
| 115–125 | 26 | 120 | 0 | 0 | 5.3 | 137.8 |
| 125–135 | 30 | 130 | 1 | 30 | 4.7 | 141 |
| 135–145 | 12 | 140 | 2 | 24 | 14.7 | 176.4 |
| 145–155 | 10 | 150 | 3 | 30 | 24.7 | 247 |
| Total | 100 | | | 53 | | 1128.8 |

$\therefore$ Mean $= \bar{x} = a + \dfrac{\Sigma f_i u_i}{\Sigma f_i} \times c = 120 + \dfrac{53}{100} \times 10 = 120 + 5.3 = 125.3$

$\therefore$ Mean deviation about the mean $= \dfrac{\Sigma f_i\,|x_i - \bar{x}|}{\Sigma f_i} = \dfrac{1128.8}{100}$

$= 11.29$ **(approx.)**

**Example 9.** *Find the mean deviation about the median for the following data :*

| *Marks* | 0–10 | 10–20 | 20–30 | 30–40 | 40–50 | 50–60 |
|---|---|---|---|---|---|---|
| *Number of girls* | 8 | 10 | 10 | 16 | 4 | 2 |

**Solution.** We construct the following table. Note that 5th and 6th columns are filled in after calculating the median.

| *Marks* | *Number of girls* $f_i$ | *Cumulative Frequency (c.f.)* | *Mid-points* $x_i$ | $\|x_i - M\|$ | $f_i\|x_i - M\|$ |
|---|---|---|---|---|---|
| 0–10 | 8 | 8 | 5 | 22 | 176 |
| 10–20 | 10 | 18 | 15 | 12 | 120 |
| 20–30 | 10 | 28 | 25 | 2 | 20 |
| 30–40 | 16 | 44 | 35 | 8 | 128 |
| 40–50 | 4 | 48 | 45 | 18 | 72 |
| 50–60 | 2 | 50 | 55 | 28 | 56 |
| | 50 | | | | 572 |

The class containing $\frac{n}{2}$th or 25th observation is 20–30, which is the median class.

$$\text{Median} = M = l + \frac{\frac{n}{2} - C}{f} \times h = 20 + \frac{25 - 18}{10} \times 10 = 27$$

$$\text{M.D. (M)} = \frac{1}{n}\Sigma f_i\,|x_i - M| = \frac{572}{50} = 11.44$$

## Limitations of mean deviation

(*i*) Mean deviation from the median is not fully reliable measure of dispersion where there is a high degree of variability, as the median is not a representative central tendency measure in such a case.

(*ii*) Also, as absolute value of deviations are taken, further algebraic treatment is not possible.

(*iii*) The sum of absolute deviations from the mean is more than the sum of the absolute deviations from the median. In some cases, this is not a reliable measure.

## EXERCISE 13.1

1. Find range of the following data:
   4, 6, 7, 8, 3, 9, 15, 8, 25, 2.
2. Calculate the quartile deviation of the following observations:
   15, 20, 22, 28, 35, 27, 44, 48, 50, 55 and 60.

3. Find the mean deviation about the median of the following data :

3, 6, 11, 12, 18.

4. Find the mean deviation about the mean of the following data :

1, 3, 7, 9, 10, 12.

5. Find the mean deviation about the median of the following data :

2, 7, 9, 11, 15, 16.

6. Find the mean deviation about the mean, as well as median for the following series:

(*i*) 6, 7, 10, 12, 13, 4, 8, 12 (*ii*) 20, 28, 40, 12, 30, 15, 50.

7. Calculate the quartile deviation for the following frequency distribution.

| *Class interval* | 30 – 34 | 35 – 39 | 40 – 44 | 45 – 49 | 50 – 54 | 55 – 59 | 60 – 64 | 65 – 69 | 70 – 74 | 75 – 79 | 80 – 84 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Frequency* | 1 | 2 | 6 | 7 | 9 | 11 | 8 | 7 | 5 | 3 | 1 |

8. Find the mean deviation about the mean for the following data :

| *Size* ($x_i$) | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |
|---|---|---|---|---|---|---|---|---|
| *Frequency* ($f_i$) | 3 | 3 | 4 | 14 | 7 | 4 | 3 | 4 |

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

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

10. 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 |

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

| *Class interval* | 0–6 | 6–12 | 12–18 | 18–24 | 24–30 |
|---|---|---|---|---|---|
| *Frequency* | 4 | 5 | 3 | 6 | 2 |

## Answers

**1.** 23 **2.** 14 **3.** 4.2 **4.** $\frac{10}{3}$ **5.** 4

**6.** (*i*) M.D. ($\bar{x}$) = 2.75, M.D. (M) = 2.75 (*ii*) M.D. ($\bar{x}$) = $\frac{512}{49}$, M.D. (M) = $\frac{73}{7}$

**7.** 8.21 **8.** 2.95 (approx.) **9.** 5.1 (approx.) **10.** 10 **11.** 7

## VARIANCE AND STANDARD DEVIATION

We have already studied the mean deviation from the mean :

$$\text{M.D.} = \frac{\Sigma\,|\,x_i - \bar{x}\,|}{n},$$

where we removed negative signs from the deviations by using the absolute value (modulus).

Another way of removing the negative signs is by squaring the deviations and then finding the average.

*The mean of squared deviations is called* ***variance,*** *denoted var*($x$) or $\sigma^2$.

*Thus, for ungrouped data :*

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

where $\bar{x}$ is the mean of the given data and $n$ is the total number of observations.

Observe that

$$\sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{\Sigma(x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n}$$

$$= \frac{\Sigma x_i^2}{n} - 2\bar{x}\left(\frac{\Sigma x_i}{n}\right) + n\frac{\bar{x}^2}{n}$$

$$= \frac{\Sigma x_i^2}{n} - \bar{x}^2. \qquad \left(\because \frac{\Sigma x_i}{n} = \bar{x}\right)$$

Thus, we can use the following **short cut method** to calculate variance

$$\sigma^2 = \frac{\Sigma x_i^2}{n} - \bar{x}^2, \text{ where } \bar{x} = \frac{\Sigma x_i}{n}.$$

Sometimes, this considerably reduces the calculation work.

A large value of variance indicates greater dispersion of values around the central value (mean), while a smaller value indicates lesser spread. One defect with variance is that its units are different from units of variable $x$. Thus, when we are talking about heights of students in centimetres, variance is in *square centimetres*. It is difficult to visualise the spread of heights in terms of square centimetres. This leads us to define **standard deviation (S.D.)** or **root mean squared deviation,** which is the square root of variance. It is denoted by $\sigma$.

Hence, for ungrouped data :

Standard deviation $\sigma = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n}}$.

It can also be taken as $\sigma = \sqrt{\dfrac{\Sigma x_i^2}{n} - \bar{x}^2}$.

When values of $(x_i - \bar{x})$ are small, we use first formula; when values of $x_i$ are convenient, we use second formula.

**Variance and standard deviation for grouped data**

If the variates (observations) $x_1, x_2, \ldots, x_m$ have frequencies $f_1, f_2, \ldots, f_m$ respectively, then the variance is defined as:

$$\text{Var}(x) \text{ or } \sigma^2 = \frac{\sum_{i=1}^{m} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{m} f_i}, \text{ where } \bar{x} = \frac{\sum_{i=1}^{m} f_i x_i}{\sum_{i=1}^{m} f_i}.$$

It is convenient to remember it as

$$\sigma^2 = \frac{\Sigma f_i(x_i - \bar{x})^2}{\Sigma f_i}, \text{ where } \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}.$$

It can be shown that this is equivalent to

$$\sigma^2 = \frac{\Sigma f_i x_i^2}{\Sigma f_i} - \bar{x}^2, \text{ where } \bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}.$$

Correspondingly, the standard deviation is

$$\sigma = \sqrt{\frac{\Sigma f_i(x_i - \bar{x})^2}{\Sigma f_i}} = \sqrt{\frac{\Sigma f_i x_i^2}{\Sigma f_i} - \bar{x}^2} = \sqrt{\frac{\Sigma f_i x_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i x_i}{\Sigma f_i}\right)^2}.$$

It can be written as

$$\sigma = \frac{1}{n}\sqrt{n\Sigma f_i x_i^2 - (\Sigma f_i x_i)^2}, \text{ where } n = \Sigma f_i.$$

Use whatever form is convenient in a given situation.

If the data is *continuous* and each class in the frequency table consists not of single value but an interval, then we use the *mid-point* of each class in the above formulae to get *approximate* values of variance and standard deviation.

**Deviation method**

To reduce the calculations further, we can use an assumed mean A, and let $d_i$ be the deviation of $x_i$ from A *i.e.* $d_i = x_i - A$. Then $x_i = d_i + A$

$$\Rightarrow \quad \Sigma f_i x_i = \Sigma f_i (d_i + A) = \Sigma f_i d_i + A\Sigma f_i$$

$$\Rightarrow \quad \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{\Sigma f_i d_i}{\Sigma f_i} + A \Rightarrow \bar{x} - A = \frac{\Sigma f_i d_i}{\Sigma f_i}.$$

Now $x_i - A = (x_i - \bar{x}) + (\bar{x} - A)$

$$\Rightarrow \quad (x_i - A)^2 = (x_i - \bar{x})^2 + (\bar{x} - A)^2 + 2(x_i - \bar{x})(\bar{x} - A)$$

$$\Rightarrow \quad \Sigma f_i (x_i - A)^2 = \Sigma f_i (x_i - \bar{x})^2 + (\bar{x} - A)^2 \Sigma f_i + 2(\bar{x} - A)\Sigma f_i (x_i - \bar{x})$$

(Note that $\bar{x} - A$, $(\bar{x} - A)^2$ are constants and can be taken out of sigma)

$$\Rightarrow \quad \Sigma f_i d_i^2 = \Sigma f_i (x_i - \bar{x})^2 + \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2 \Sigma f_i + 0$$

$$\left(\text{As } \Sigma f_i (x_i - \bar{x}) = \Sigma f_i x_i - \Sigma f_i \bar{x} = \Sigma f_i \left(\frac{\Sigma f_i x_i}{\Sigma f_i} - \bar{x}\right) = \Sigma f_i (\bar{x} - \bar{x}) = 0\right)$$

$$\Rightarrow \quad \frac{\Sigma f_i d_i^2}{\Sigma f_i} = \frac{\Sigma f_i (x_i - \bar{x})^2}{\Sigma f_i} + \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2.$$

$$\therefore \quad \sigma^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{\Sigma f_i} = \frac{\Sigma f_i d_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2.$$

For ungrouped data, it becomes

$$\sigma^2 = \frac{\Sigma d_i^2}{n} - \left(\frac{\Sigma d_i}{n}\right)^2.$$

**Step deviation method**

If in a continuous grouped data, the classes are of uniform size, say $c$, the calculations can be further simplified.

Putting $t_i = \frac{d_i}{c}$, we get

$$\sigma^2 = \frac{\Sigma f_i d_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2 = \frac{\Sigma f_i (t_i c)^2}{\Sigma f_i} - \left(\frac{\Sigma f_i t_i c}{\Sigma f_i}\right)^2$$

$$= c^2\left[\frac{\Sigma f_i t_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i t_i}{\Sigma f_i}\right)^2\right]$$

$$\Rightarrow \quad \sigma = c\sqrt{\frac{\Sigma f_i t_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i t_i}{\Sigma f_i}\right)^2}, \text{ where } t_i = \frac{d_i}{c} = \frac{x_i - A}{c}.$$

## ILLUSTRATIVE EXAMPLES

**Example 1.** *Find the mean, variance and standard deviation for the following data :*
*6, 10, 7, 13, 4, 12, 8, 12*

**Solution.** Here the number of observations $n = 8$.

$$\text{Mean} = \overline{x} = \frac{\Sigma x_i}{n} = \frac{6 + 10 + 7 + 13 + 4 + 12 + 8 + 12}{8} = \frac{72}{8} = 9$$

The respective $x_i - \overline{x}$ are

$$6 - 9, 10 - 9, 7 - 9, 13 - 9, 4 - 9, 12 - 9, 8 - 9, 12 - 9$$

*i.e.* $-3, 1, -2, 4, -5, 3, -1, 3$

$$\therefore \quad \Sigma(x_i - \overline{x})^2 = (-3)^2 + 1^2 + (-2)^2 + 4^2 + (-5)^2 + 3^2 + (-1)^2 + 3^2$$

$$= 9 + 1 + 4 + 16 + 25 + 9 + 1 + 9 = 74$$

$$\therefore \quad \text{Variance} = \frac{\Sigma(x_i - \overline{x})^2}{n} = \frac{74}{8} = 9.25$$

Standard deviation = $\sqrt{9.25}$ = 3.04 (approx.)

**Example 2.** *Find the mean and the variance of first 10 multiples of 3.*

**Solution.** The first 10 multiples of 3 are : 3, 6, 9, 12, 15, 18, 21, 24, 27, 30.

Here number of observations $n = 10$.

$$\text{Sum of all observations} = 3 + 6 + 9 + \ldots + 30$$

$$= 3(1 + 2 + 3 + \ldots + 10)$$

$$= 3 \times \frac{10(10 + 1)}{2} = 15 \times 11 = 165$$

$$\therefore \quad \text{Mean} = \overline{x} = \frac{\Sigma x_i}{n} = \frac{165}{10} = 16.5$$

$$\therefore \quad \Sigma x_i^2 = 3^2 + 6^2 + 9^2 + \ldots + 30^2$$

$$= 3^2(1^2 + 2^2 + 3^2 + \ldots + 10^2)$$

$$= 9 \times \frac{10(10 + 1)(2 \times 10 + 1)}{6} = 15 \times 11 \times 21 = 3465.$$

$$\therefore \quad \text{Variance } \sigma^2 = \frac{\Sigma x_i^2}{n} - (\overline{x})^2 = \frac{3465}{10} - (16.5)^2$$

$$= 346.5 - 272.25 = 74.25$$

*Alternatively*

First 10 multiples of 3 are : 3, 6, 9, 12, 15, 18, 21, 24, 27, 30.

Taking assumed mean A = 15, we construct the table as under :

| $x_i$ | $d_i = x_i - A = x_i - 15$ | $d_i^2$ |
|---|---|---|
| 3 | −12 | 144 |
| 6 | −9 | 81 |
| 9 | −6 | 36 |
| 12 | −3 | 9 |
| 15 | 0 | 0 |
| 18 | 3 | 9 |
| 21 | 6 | 36 |
| 24 | 9 | 81 |
| 27 | 12 | 144 |
| 30 | 15 | 225 |
| Total | 15 | 765 |

Here $n = 10$, $\Sigma d_i = 15$ and $\Sigma d_i^2 = 765$.

$$\therefore \quad \text{Mean} = A + \frac{\Sigma d_i}{n} = 15 + \frac{15}{10} = 15 + 1.5 = 16.5$$

$$\text{Variance} = \frac{\Sigma d_i^2}{n} - \left(\frac{\Sigma d_i}{n}\right)^2 = \frac{765}{10} - \left(\frac{15}{10}\right)^2 = 76.5 - (1.5)^2$$

$$= 76.5 - 2.25 = 74.25$$

**Example 3.** *Find the mean, standard deviation and variance of the first n natural numbers.*

**Solution.** The given numbers are 1, 2, 3, …, $n$.

$$\text{Mean} = \bar{x} = \frac{\Sigma n}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}.$$

$$\text{Variance } \sigma^2 = \frac{\Sigma x_i^2}{n} - \bar{x}^2 = \frac{\Sigma n^2}{n} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4}$$

$$= (n+1)\left[\frac{2n+1}{6} - \frac{n+1}{4}\right] = (n+1)\left(\frac{n-1}{12}\right) = \frac{n^2-1}{12}.$$

$$\therefore \text{ Standard deviation } \sigma = \sqrt{\frac{n^2-1}{12}}.$$

**Example 4.** *Find the mean, 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.** We construct the following table. Note that 4th, 5th and 6th columns are filled in after calculating the mean.

| $x_i$ | $f_i$ | $f_i x_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $f_i(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 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 |
| Total | 30 | 420 | | | 1374 |

Here $n = \Sigma f_i = 30$, $\Sigma f_i x_i = 420$.

$\therefore$ Mean $= \bar{x} = \dfrac{\Sigma f_i x_i}{n} = \dfrac{420}{30} = 14.$

$\therefore$ Variance $\sigma^2 = \dfrac{1}{n}\Sigma f_i(x_i - \bar{x})^2 = \dfrac{1}{30} \times 1374 = 45.8$

$\therefore$ Standard deviation $\sigma = \sqrt{45.8} = 6.77$ (approx.)

**Example 5.** *Calculate the standard deviation for the following data :*

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

**Solution.** To simplify calculations, we will use

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

| $x_i$ | $f_i$ | $f_i x_i$ | $x_i^2$ | $f_i x_i^2$ |
|---|---|---|---|---|
| 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 |
| Total | 48 | 614 | | 9652 |

Here, $n = \Sigma f_i = 48$, $\Sigma f_i x_i = 614$

$$\sigma = \frac{1}{n}\sqrt{n\Sigma f_i x_i^2 - (\Sigma f_i x_i)^2} = \frac{1}{48}\sqrt{48 \times 9652 - (614)^2}$$

$$= \frac{1}{48}\sqrt{463296 - 376996} = \frac{1}{48}\sqrt{86300} = \frac{293.76}{48} = 6.12 \text{ (approx.)}$$

Therefore, the standard deviation $\sigma = 6.12$ (approx.)

**Example 6.** *Calculate the mean, variance and standard deviation of the following data :*

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

**Solution.** Taking assumed mean A = 65, we construct the following table :

| Classes | Mid-points $x_i$ | Frequency $f_i$ | $d_i = x_i - 65$ | $d_i^2$ | $f_i d_i$ | $f_i d_i^2$ |
|---|---|---|---|---|---|---|
| 30 – 40 | 35 | 3 | – 30 | 900 | –90 | 2700 |
| 40 – 50 | 45 | 7 | –20 | 400 | –140 | 2800 |
| 50 – 60 | 55 | 12 | –10 | 100 | –120 | 1200 |
| 60 – 70 | 65 | 15 | 0 | 0 | 0 | 0 |
| 70 – 80 | 75 | 8 | 10 | 100 | 80 | 800 |
| 80 – 90 | 85 | 3 | 20 | 400 | 60 | 1200 |
| 90 – 100 | 95 | 2 | 30 | 900 | 60 | 1800 |
| Total | | 50 | | | –150 | 10500 |

Here, $\Sigma f_i = 50$, $\Sigma f_i d_i = -150$ and $\Sigma f_i d_i^2 = 10500$.

$$\therefore \quad \text{Mean} = A + \frac{\Sigma f_i di}{\Sigma f_i} = 65 + \left(\frac{-150}{50}\right) = 65 - 3 = 62.$$

$$\text{Variance } \sigma^2 = \frac{\Sigma f_i d_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i d_i}{\Sigma f_i}\right)^2 = \frac{10500}{50} - \left(\frac{-150}{50}\right)^2 = 210 - 9 = 201.$$

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

**Example 7.** *Find the mean, variance and standard deviation for the following data :*

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

**Solution.** Since the classes are of uniform width 30, we use the step deviation method with assumed mean A = 105 and $t_i = \frac{x_i - A}{c} = \frac{x_i - 105}{30}$.

We construct the table as under :

| Classes | Mid-points $x_i$ | Frequency $f_i$ | $t_i = \frac{x_i - 105}{30}$ | $t_i^2$ | $f_i t_i$ | $f_i t_i^2$ |
|---|---|---|---|---|---|---|
| 0 – 30 | 15 | 2 | –3 | 9 | –6 | 18 |
| 30 – 60 | 45 | 3 | –2 | 4 | –6 | 12 |
| 60 – 90 | 75 | 5 | –1 | 1 | –5 | 5 |
| 90 – 120 | 105 | 10 | 0 | 0 | 0 | 0 |
| 120 – 150 | 135 | 3 | 1 | 1 | 3 | 3 |
| 150 – 180 | 165 | 5 | 2 | 4 | 10 | 20 |
| 180 – 210 | 195 | 2 | 3 | 9 | 6 | 18 |
| Total | | 30 | | | 2 | 76 |

Here, $\Sigma f_i = 30$, $\Sigma f_i t_i = 2$ and $\Sigma f_i t_i^2 = 76$.

$$\therefore \quad \text{Mean} = A + c \times \frac{\Sigma f_i t_i}{\Sigma f_i} = 105 + 30 \times \frac{2}{30} = 105 + 2 = 107.$$

$$\text{Variance} = c^2 \left[\frac{\Sigma f_i t_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i t_i}{\Sigma f_i}\right)^2\right] = (30)^2 \left[\frac{76}{30} - \left(\frac{2}{30}\right)^2\right] = 76 \times 30 - 4 = 2276.$$

$$\text{Standard deviation} = \sqrt{2276} = 47.71 \text{ (approx.)}$$

**Example 8.** *Calculate the mean and standard deviation for the following data :*

| *Wages per day upto (in ₹)* | 100 | 200 | 300 | 400 | 500 | 600 | 700 |
|---|---|---|---|---|---|---|---|
| *Number of workers* | 9 | 26 | 58 | 81 | 121 | 139 | 140 |

**Solution.** Since the data is given in the form of cumulative frequency, first we convert it into continuous frequency distribution as under :

| *Wages per day (in ₹)* | 0–100 | 100–200 | 200–300 | 300–400 | 400–500 | 500–600 | 600–700 |
|---|---|---|---|---|---|---|---|
| *No. of workers* | 9 | 17 | 32 | 23 | 40 | 18 | 1 |

As the classes are of uniform width 100, we use step deviation method with assumed mean $A = 350$ and $t_i = \frac{x_i - A}{c} = \frac{x_i - 350}{100}$.

We construct the table as under :

| *Classes* | *Class mark* $x_i$ | *Frequency* $f_i$ | $t_i = \frac{x_i - 350}{100}$ | $t_i^2$ | $f_i t_i$ | $f_i t_i^2$ |
|---|---|---|---|---|---|---|
| 0–100 | 50 | 9 | –3 | 9 | –27 | 81 |
| 100–200 | 150 | 17 | –2 | 4 | –34 | 68 |
| 200–300 | 250 | 32 | –1 | 1 | –32 | 32 |
| 300–400 | 350 | 23 | 0 | 0 | 0 | 0 |
| 400–500 | 450 | 40 | 1 | 1 | 40 | 40 |
| 500–600 | 550 | 18 | 2 | 4 | 36 | 72 |
| 600–700 | 650 | 1 | 3 | 9 | 3 | 9 |
| Total | | 140 | | | –14 | 302 |

$$\therefore \quad \text{Mean} = A + c \times \frac{\Sigma f_i t_i}{\Sigma f_i} = 350 + 100\left(\frac{-14}{140}\right) = 350 - 10 = 340.$$

$$\text{Variance} = c^2 \left[\frac{\Sigma f_i t_i^2}{\Sigma f_i} - \left(\frac{\Sigma f_i t_i}{\Sigma f_i}\right)^2\right] = (100)^2 \left[\frac{302}{140} - \left(\frac{-14}{140}\right)^2\right]$$

$$= (100)^2\,[2.1571 - 0.01] = 10000 \times 2.1471 = 21471$$

$\therefore$ Standard deviation $= \sqrt{21471} = 146.5$

$\therefore$ Mean = ₹340 and standard deviation = ₹146.5

**Example 9.** *The mean and variance of 7 observations are 8 and 16 respectively. If five of the observations are 2, 4, 10, 12 and 14, find the remaining two observations.*

**Solution.** Let the remaining two observations be $x$ and $y$.

Given mean $\bar{x} = 8 \Rightarrow \dfrac{2+4+10+12+14+x+y}{7} = 8$

$\Rightarrow \quad 42 + x + y = 56$

$\Rightarrow \quad x + y = 14 \qquad \ldots(i)$

Also variance $\sigma^2 = 16 \Rightarrow \dfrac{\Sigma x_i^2}{n} - (\bar{x})^2 = 16$

$\Rightarrow \quad \dfrac{2^2+4^2+10^2+12^2+14^2+x^2+y^2}{7} - (8)^2 = 16$

$\Rightarrow \quad 4 + 16 + 100 + 144 + 196 + x^2 + y^2 = 7(16 + 64)$

$\Rightarrow \quad 460 + x^2 + y^2 = 560$

$\Rightarrow \quad x^2 + y^2 = 100$

$\Rightarrow \quad x^2 + (14 - x)^2 = 100 \qquad$ (using $(i)$)

$\Rightarrow \quad 2x^2 - 28x + 196 - 100 = 0$

$\Rightarrow \quad x^2 - 14x + 48 = 0$

$\Rightarrow \quad (x - 6)(x - 8) = 0 \Rightarrow x = 6, 8.$

When $x = 6$, $y = 14 - 6 = 8$; when $x = 8$, $y = 14 - 8 = 6$.

Hence, the remaining two observations are 6 and 8.

**Example 10.** *The mean and standard deviation of 6 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.*

**Solution.** Let $x_1, x_2, \ldots, x_6$ be the six given observations.

Then $\bar{x} = 8$ and $\sigma = 4$.

Now $\bar{x} = \dfrac{\Sigma x_i}{n} \Rightarrow 8 = \dfrac{x_i + x_2 + \ldots + x_6}{6}$

$\Rightarrow \quad x_1 + x_2 + \ldots + x_6 = 48 \qquad \ldots(i)$

Also $\sigma^2 = \dfrac{\Sigma x_i^2}{n} - (\bar{x})^2 \Rightarrow 4^2 = \dfrac{x_1^2 + x_2^2 + \ldots + x_6^2}{6} - (8)^2$

$\Rightarrow \quad {x_1}^2 + {x_2}^2 + \ldots + {x_6}^2 = 6 \times (16 + 64) = 480 \qquad \ldots(ii)$

As each observation is multiplied by 3, new observations are

$3x_1, 3x_2, \ldots, 3x_6$.

New mean $\bar{X} = \dfrac{3x_1 + 3x_2 + \ldots + 3x_6}{6} = \dfrac{3(x_1 + x_2 + \ldots + x_6)}{6}$

$= \dfrac{3 \times 48}{6} \qquad$ (using $(i)$)

$= 24 \qquad \ldots(iii)$

Let $\sigma_1$ be the new standard deviation, then

$$\sigma_1^2 = \frac{(3x_1)^2 + (3x_2)^2 + \dots + (3x_6)^2}{6} - (\bar{X})^2$$

$$= \frac{9(x_1^2 + x_2^2 + \dots + x_6^2)}{6} - (24)^2 \qquad \text{(using } (iii))$$

$$= \frac{9 \times 480}{6} - 576 \qquad \text{(using } (ii))$$

$$= 720 - 576 = 144$$

$\Rightarrow \quad \sigma_1 = 12.$

**Example 11.** *Given that $\bar{x}$ is the mean and $\sigma^2$ is the variance of n observations $x_1, x_2, x_3, \dots, x_n$. Prove that the mean and variance of the observations $ax_1, ax_2, ax_3, \dots, ax_n$ $(a \neq 0)$ are $a\bar{x}$ and $a^2\sigma^2$.*

**Solution.** Given observations are $x_1, x_2, x_3, \dots, x_n$. The mean is $\bar{x}$ and variance $= \sigma^2$.

$$\therefore \quad \bar{x} = \frac{\Sigma x_i}{n} \quad \dots(i) \qquad \text{and} \qquad \sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{n} \quad \dots(ii)$$

The new observations are $ax_1, ax_2, ax_3, \dots, ax_n$ $(a \neq 0)$

*i.e.* $\quad y_i = ax_i, i = 1, 2, 3, \dots, n.$

Let $\bar{y}$ be the mean of these new observations, then

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{\Sigma ax_i}{n} = a \times \frac{\Sigma x_i}{n} = a\bar{x} \qquad \text{(using } (i))$$

$$\text{New variance} = \sigma_1^2 = \frac{\Sigma (y_i - \bar{y})^2}{n} = \frac{\Sigma (ax_i - a\bar{x})^2}{n}$$

$$= \frac{\Sigma (a(x_i - \bar{x}))^2}{n} = a^2 \times \frac{\Sigma (x_i - \bar{x})^2}{n}$$

$$= a^2\sigma^2 \qquad \text{(using } (ii))$$

Hence, new mean $= a\bar{x}$ and new variance $= a^2\sigma^2$.

**Example 12.** *If each of the observations $x_1, x_2, \dots, x_n$ is increased by an amount 'a' where 'a' is a positive or negative number, show that the variance and standard deviation remain the same.*

**Solution.** Let $\bar{x}$ be the mean of $x_1, x_2, \dots, x_n$.

Variance $\sigma^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2.$

If $a$ is added to each observation, then new observations are

$y_i = x_i + a$, for $i = 1, 2, \dots, n$

Let $\bar{y}$ be the mean of these observations, then

$$\bar{y} = \frac{1}{n}\Sigma y_i = \frac{1}{n}\Sigma (x_i + a) = \frac{1}{n}[\Sigma x_i + \Sigma a]$$

$$= \frac{1}{n}[\Sigma x_i + na] = \left(\frac{1}{n}\Sigma x_i\right) + a = \bar{x} + a$$

*i.e.* $\quad \bar{y} = \bar{x} + a.$

$\therefore$ The variance of new observations

$$\sigma_1^2 = \frac{1}{n}\Sigma(y_i - \bar{y})^2 = \frac{1}{n}\Sigma[(x_i + a) - (\bar{x} + a)]^2$$

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

$\Rightarrow \quad \sigma_1 = \sigma.$

Thus, while the mean changes to $\bar{x} + a$, the variance and standard deviation remain the same.

**Example 13.** *The mean and standard deviation of 100 observations were calculated as 40 and 5.1 respectively by a student who mistook one observation as 50 instead of 40. What are the correct mean and standard deviation?*

**Solution.** Given that $n = 100$, incorrect mean $\bar{x} = 40$, incorrect S.D. $(\sigma) = 5.1$

As $\bar{x} = \frac{\Sigma x_i}{n}$,

$$40 = \frac{\Sigma x_i}{100} \Rightarrow \Sigma x_i = 4000$$

$\Rightarrow$ incorrect sum of observations = 4000

$\Rightarrow$ correct sum of observations = 4000 – 50 + 40 = 3990

So correct mean = $\frac{3990}{100} = 39.9$

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

Using incorrect value,

$$5.1 = \sqrt{\frac{1}{100}\Sigma x_i^2 - (40)^2}$$

$\Rightarrow \quad 26.01 = \left[\frac{1}{100}\Sigma x_i^2 - 1600\right] \Rightarrow \Sigma x_i^2 = 2601 + 160000 = 162601$

$\Rightarrow$ incorrect $\Sigma x_i^2 = 162601$

$\Rightarrow$ correct $\Sigma x_i^2 = 162601 - (50)^2 + (40)^2 = 162601 - 2500 + 1600 = 161701$

$\therefore \quad$ Correct $\sigma = \sqrt{\frac{1}{100}\text{correct } \Sigma x_i^2 - (\text{correct } \bar{x})^2}$

$$= \sqrt{\frac{1}{100}(161701) - (39.9)^2} = \sqrt{1617.01 - 1592.01}$$

$$= \sqrt{25} = 5$$

Hence, correct mean is 39.9 and correct standard deviation is 5.

**Example 14.** *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 the three observations were incorrect, which were recorded as 21, 21 and 18. Find the mean and the standard deviation if the incorrect observations are omitted.*

**Solution.** Given $n = 100$, mean = 20 and standard deviation = 3

*i.e.* variance = 9.

Now mean = 20 $\Rightarrow \frac{\sum_{i=1}^{100} x_i}{100} = 20 \Rightarrow \sum_{i=1}^{100} x_i = 20 \times 100 = 2000$

and variance = 9 $\Rightarrow \dfrac{\sum_{i=1}^{100} x_i^2}{100} - (20)^2 = 9$

$$\Rightarrow \sum_{i=1}^{100} x_i^2 = 100(9 + 400) = 40900.$$

When the three incorrect observations 21, 21 and 18 are omitted from the data, then 97 observations are left.

$\therefore$ Correct $\sum_{i=1}^{97} x_i = 2000 - 21 - 21 - 18 = 1940.$

$$\therefore \text{ Correct mean} = \frac{\text{correct} \sum_{i=1}^{97} x_i}{97} = \frac{1940}{97} = 20.$$

$$\text{Correct } \sum_{i=1}^{97} x_i^2 = 40900 - 21^2 - 21^2 - 18^2$$

$$= 40900 - 441 - 441 - 324 = 39694.$$

$$\therefore \text{ Correct } \sigma^2 = \frac{\text{correct} \sum_{i=1}^{97} x_i^2}{97} - (\text{correct mean})^2$$

$$= \frac{39694}{97} - (20)^2 = 409.22 - 400 = 9.22$$

$\therefore$ Correct standard deviation = $\sqrt{9.22} = 3.036$

## EXERCISE 13.2

1. Find the variance and the standard deviation for the following data:

   3, 6, 11, 12, 18.

2. Find the variance and the standard deviation for the following data:

   1, 3, 7, 9, 10, 12.

3. Find the mean, variance and standard deviation of the following marks scored by 10 students:

   45, 70, 62, 60, 50, 48, 67, 34, 65, 58.

4. Find the mean and standard deviation for the following data:

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

5. Find the mean, variance and standard deviation for the following data by 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 |

6. Find the mean, variance and standard deviation for the following frequency distribution:

| *Classes* | 0–10 | 10–20 | 20–30 | 30–40 | 40–50 |
|---|---|---|---|---|---|
| *Frequency* | 5 | 8 | 15 | 16 | 6 |

7. Calculate the mean, variance and standard deviation for the following frequency distribution by using step deviation method :

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

8. Find the mean, variance and standard deviation for the following data by using step deviation method :

| Classes | 10–20 | 20–30 | 30–40 | 40–50 | 50–60 | 60–70 | 70–80 | 80–90 | 90–100 |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | 3 | 4 | 7 | 7 | 15 | 9 | 6 | 6 | 3 |

9. The heights, to the nearest cm, of 30 men are 159, 170, 174, 173, 175, 160, 161, 164, 163, 165, 164, 171, 162, 170, 177, 185, 181, 180, 175, 165, 186, 174, 168, 168, 176, 176, 165, 175, 167 and 180. Using class intervals 155 – 160, 160 – 165, ... draw up a grouped frequency distribution and use this to estimate the arithmetic mean and the standard deviation.

10. The mean of 5 observations is 6 and the standard deviation is 2. If the three observations are 5, 7 and 9, find the other two observations.

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

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

13. While calculating the mean and variance of 10 readings, a student wrongly uses the reading 52 for the correct reading 25. He obtained the mean and variance as 45 and 16 respectively. Find the correct mean and variance.

14. 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 correct standard deviation in each of the following cases :

(*i*) If the wrong observation is omitted.

(*ii*) If it is replaced by 12.

## Answers

**1.** 26.8; $\sqrt{26.8}$ **2.** 15; $\sqrt{15}$ **3.** 55.9; 115.89; 10.77 **4.** 19; 6.59

**5.** 64; 2.86; 1.69 **6.** 27; 132; 11.49 **7.** 62; 201; 14.18 **8.** 56; 422.33; 20.55

**9.** 172; 7.9 **10.** 3 and 6 **11.** 4 and 9 **12.** 20

**13.** 42.3; 43.81 **14.** (*i*) 10.11; 1.997 (*ii*) 10.2; 1.99