(ii) A comparison between collection of data is possible.

(iii) They allow for additional calculations and inferences.

A set of data can be completely describe by two numerical quantities. They are:

(i) Measure of central tendency.

(ii) Measure of dispersion.

Measure of Central tendency

Measure of central tendency is a numerical value which indicates the concentration of the data or how the data tend to build-up about that value. The three methods of measure of central tendency are:

a. Mean or Arithmetic mean or Average

b. Median

c. Mode

(a) Mean: Mean of a set of data is the arithmetic average of the set of data and is given by the sum of the values of the datas divided by total number of observations. Mean is denoted by $\overline{X}$ (read as x bar) representing the mean of x values of the data. Mean is the most commonly measure of central tendency.

Methods of Calculation of mean

(i) Raw data technique.

(ii) Ungrouped data technique.

(iii) Grouped data technique.

(i) Raw data technique: Raw data technique is used when the number observations in the data are small. In this case mean is given by the formula.

$$\overline{X} = \left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \left[\frac{X_1 + X_2 + X_3 + ... + X_n}{n}\right]$$

Where, $X_1, X_2, ...X_n$ are individual observed data and n is the number of observations.

Illustration 6.1: The diameters of six cylinders in mms are as follows:25.2, 26.5, 20.1, 13.9, 27.2 and 22.4. Compute the mean diameters of the Cylinders.

Solution:

Number of Observations = $n = 6$

$X_1 = 25.2, X_2 = 26.5, X_3 = 20.1, X_4 = 13.9, X_5 = 27.2$ *and* $X_6 = 22.4$

We have,

$$\bar{X} = \left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \left[\frac{\sum_{i=1}^{6} X_i}{6}\right] = \left[\frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}\right]$$

$$= \frac{[25.2 + 26.5 + 20.1 + 13.9 + 27.2 + 22.4]}{6} = \left[\frac{135.3}{6}\right] = 22.55 \text{ mm}$$

(ii) Ungrouped data technique: In this case the data will be un the form of frequency distribution and ungrouped. In such cases the mean is given by the formula.

$$\bar{X} = \left[\frac{\sum_{i=1}^{N} f_i X_i}{\sum f \text{ or } \sum n}\right] = \left[\frac{f_1 x_1 + f_2 x_2 + \ldots\ldots\ldots\ldots f_n x_n}{\sum f \text{ or } \sum n}\right]$$

Where, $\sum f$ or $\sum n$ is the total number of observations in the data.

Illustration 6.2: Table 6.1 shows the frequency distribution of ungrouped data. Compute the mean of the data.

Table 6.1 Frequency distribution data

		Given	Computed
Sl.No.	X_i	f_i	$f_i x_i$
1	15	10	150
2	16	12	192
3	17	22	374
4	18	09	162
5	19	07	133
Total number observations		$\sum f$ or $\sum n = 60$	$\sum f_i x_i = 1011$

Solution:

$$\bar{X} = \frac{\sum f_i x_i}{n} = \left(\frac{1011}{60}\right) = 16.85$$

(iii) Grouped data technique

When the data have been grouped into cells or class and frequency distribution have been drawn, then the **mean** of such data can been computed using the formula. Datas are grouped into classes or cells, when the number of observation and categories are large.

$$\overline{X} = \left[\frac{\sum_{j=1}^{N} f_j X_j}{\sum f \text{ or } \sum n} \right]$$

Where, f_j is the frequency of the j th class.

X_j is the cell midpoint of the j th class

N is the number of classes.

Illustration 6.3: Table 6.2 shows the grouped frequency distribution of life of 320 tyres in 1000 km. Compute the mean of the data.

Table 6.2: Frequency distribution of life of 320 tyres

	Given		Computed	
Cell Number	*Cell Boundaries*	*frequency*	*Cell Midpoint*	$f_j X_j$
1	23.5 – 26.5	4	25.0	100
2	26.5 – 29.5	36	28.0	1008
3	29.5 – 32.5	51	31.0	1581
4	32.5 – 35.5	63	34.0	2142
5	35.5 – 38.5	58	37.0	2146
6	38.5 – 41.5	52	40.0	2080
7	41.5 – 44.5	34	43.0	1462
8	44.5 – 47.5	16	46.0	736
9	47.5 – 50.5	6	49.0	294
Total	$\sum f$ or $\sum n$	320	$\sum f_j X_j$	11549

Solution :

$n = 9$ (*number of cells or classes*)

$\sum f \text{ or } \sum n = 320$

$$X_i = \left(\frac{23.5 + 26.5}{2}\right) = \left(\frac{50}{2}\right) = 25$$

$$Class\ width\ i = (26.5 - 23.5) = 3$$

$$\sum f_j X_j = 11549$$

$$We\ have\ \overline{X} = \frac{\sum f_j X_j}{\sum f\ or \sum n}$$

$$= \left(\frac{11549}{320}\right)$$

$$= 36.09\ (in\ 1000\ km)$$

$$\therefore\ \boxed{\overline{X} = 36.09 \times 10^3\ km}$$

The above method is quite laborious and need a calculator or a computer and there is every possibility of committing error. Another method of computation of mean of a grouped data is by deviation method.

In this method one of the class midpoint is assumed as the mean of the given data and the corresponding deviation of class or cell number from cell number of the assumed mean is found out. The product of the class mid-point and the Corresponding deviation is found out and the mean of the data is computed using the formula.

$$\overline{X} = \overline{X_0} + i\left(\frac{\sum_{j=1}^{m} f_j d_j}{m}\right)$$

Where, $\overline{X_0}$ is the assumed mean (class midpoint)

i is the class width.

f_j is the frequency of the jth class.

d_j is the deviation of the jth class from the assumed mean class.

n is the total number of observation in the data.

m is the total number of cells or classes.

Illustration 6.4: Table 6.3 below shows the grouped frequency distribution of life of 320 tyres in 1000 kms. Compute the mean of the life of tyres by the method of deviation.

Table 6.3

Given			*Computed*			
Cell Number	*Cell Interval*	*frequency* f_i	*Cell midpoint* X_i	*Deviation* d_i	$f_i d_i$	
1	23.5 – 26.5	4	25.0	–4	–16.0	
2	26.5 – 29.5	36	28.0	–3	–108.0	
3	29.5 – 32.5	51	31.0	–2	–102.0	
4	32.5 – 35.5	63	34.0	–1	–63.0	(–289)
5	35.5 – 38.5	58	37.0 = $\overline{X_s}$	0	0.0	
6	38.5 – 41.5	52	40.0	+1	52.0	
7	41.5 – 44.5	34	43.0	+2	68.0	
8	44.5 – 47.5	16	46.0	+3	48.0	
9	47.5 – 50.5	6	49.0	+4	24.0	(+192)
Total		$\sum n = 320$			$\sum f_i d_i = -97$	

Mid-Point of any class can be assumed as mean $\overline{X_s}$ of the given data.

Here the mid-point of fifth class i.e. 37.0 is the assumed mean. The deviation of all classes lower than the assumed class will be negative and similarly the deviation of all classes higher than the assumed class will be positive.

For Example deviation of the assumed class from the assumed class by class will be zero.

Similarly deviation of the immediate deviations next lower class will be -1 and subsequent lower classes will be –2, –3 and so on. On the same line the deviation by class of the immediate next higher class to the assumed class will be +1 and subsequent higher classes deviations will be +2, +3 and so on.

The product of X_i and d_i is computed for each class and then $\sum f_i X_i$ is computed taking sign of $X_i d_i$ into account.

Class width i is the difference between upper and lower boundaries of any class or the difference between any two successive class mid-point.

i.e., $i = (26.5 - 23.5) = 3$ or $(28.0 - 25.0) = 3$

Then the mean is Computed using the formula

Illustration 6.4: Table 6.3 below shows the grouped frequency distribution of life of 320 tyres in 1000 kms. Compute the mean of the life of tyres by the method of deviation.

Table 6.3

Given			*Computed*			
Cell Number	*Cell Interval*	*frequency* f_j	*Cell midpoint* X_j	*Deviation* d_j	$f_j d_j$	
1	23.5 – 26.5	4	25.0	–4	–16.0	
2	26.5 – 29.5	36	28.0	–3	–108.0	
3	29.5 – 32.5	51	31.0	–2	–102.0	
4	32.5 – 35.5	63	34.0	–1	–63.0	(–289)
5	35.5 – 38.5	58	37.0 = $\overline{X_0}$	0	0.0	
6	38.5 – 41.5	52	40.0	+1	52.0	
7	41.5 – 44.5	34	43.0	+2	68.0	
8	44.5 – 47.5	16	46.0	+3	48.0	
9	47.5 – 50.5	6	49.0	+4	24.0	(+192)
Total		$\sum n = 320$			$\sum f_j d_j = -97$	

Mid-Point of any class can be assumed as mean $\overline{X_0}$ of the given data.

Here the mid-point of fifth class i.e. 37.0 is the assumed mean. The deviation of all classes lower than the assumed class will be negative and similarly the deviation of all classes higher than the assumed class will be positive.

For Example deviation of the assumed class from the assumed class by class will be zero.

Similarly deviation of the immediate deviations next lower class will be -1 and subsequent lower classes will be –2, –3 and so on. On the same line the deviation by class of the immediate next higher class to the assumed class will be +1 and subsequent higher classes deviations will be +2, +3 and so on.

The product of X_j and d_j is computed for each class and then $\sum f_j X_j$ is computed taking sign of $X_j d_j$ into account.

Class width i is the difference between upper and lower boundaries of any class or the difference between any two successive class mid-point.

i.e., $i = (26.5 - 23.5) = 3$ or $(28.0 - 25.0) = 3$

Then the mean is Computed using the formula

$$\bar{X} = \bar{X}_0 + i\left(\frac{\sum_{i=1}^{n} f_i X_i}{\sum n}\right) = 37.0 + 3\left(\frac{-97}{320}\right) = 36.09 (in\ 1000\ km)$$

$$\therefore \boxed{\bar{X} = 36.09 \times 10^3\ kms}$$

(b) Median: Median is another method of measure of central tendency.

Median of a set of numbers or numerical data *arranged in an order* is the middle value of the set of numbers. (Arranged in an order means arranged either in ascending or descending order of its magnitude. Note the data should be arranged and then only the median should be found out)

For Example the median of a set of number 7,11,8, 16,12 .The given data is a set of raw data, and it should be arranged in either ascending or descending order.

i.e., 7, 8,11,12,16 or 16,12,11,8,7 and the middle value of the arranged data is 11 and hence 11 is the median. In this case the number of observation is 5 which is odd.

When the number of observations are even such as 6,8 and so on, then in such cases median will be the arithmetic mean of two middle values of the arranged data

Illustration 6.5: Compute the median of the set of numbers 17, 10, 9, 16,11,14. Here the number of observations are n = 6.

Arranging them in either ascending or descending order we have

7,10,11,14,16 or 16,14,11,10,9.7

And the two middle values are 10 and 11, and the arithmetic mean of these two numbers are $\frac{10+11}{2} = 10.5$.

For *grouped data* the median is Computed using the formula

$$M_d = L_m + i\left(\frac{\frac{n}{2} - Cf_m}{f_m}\right)$$

Where, L_m = Lower boundary of the median class

i = Class width

n = Total number of observation in the data

Cf_m = Cumulative frequency of all classes lower than the median class

f_m = Frequency of the median class.

Median class is the middle class of the grouped data i.e. when the number of classes $m=7$, then 4th class is the median class and when $m = 9$, 5th class will be the median class.

Illustration 6.6: Table 6.4 show the grouped frequency distribution of life of 320 tyres in 1000 kms. Compute the median of the data.

Table 6.4 Frequency distribution of life of 320 tyres

Given			Computed
Cell Number m	*Cell Interval*	*frequency* f_i	*Cumulative frequency* $\sum fi$
1	23.5 – 26.5	4	4
2	26.5 – 29.5	36	40
3	29.5 – 32.5	51	91
4	32.5 – 35.5	63	164
5	35.5 – 38.5	58	22
6	38.5 – 41.5	52	24
7	41.5 – 44.5	34	298
8	44.5 – 47.5	16	314
9	47.5 – 50.5	6	320
Total		$n = 320$	

Solution :

It can be observed from the given frequency distribution that the number of classes $m = 9$, and hence 5th class with the Class internal (35.5 – 38.5) is taken as the median class

Here, $L_m = 35.5$; $i = (38.5 - 35.5) = 3$; $Cf_m = 164$; $n = 320$; $f_m = 58$

$$\therefore M_d = L_m + i\left(\frac{\frac{n}{2} - cf_m}{f_m}\right)$$

$$= 35.5 + 3\left(\frac{\frac{320}{3} - 164}{58}\right) = 35.5 + 3(-0.069) = 35.293\ (in\ 1000\ km)$$

$$\therefore \boxed{M_d = 35.29 \times 10^3\ km}$$

Suppose the number of cells **m** are even, say 8. To compute median of such grouped data, the following procedure is adopted to find the median class and then the same formula is used to find the median.

To find the median class when the number of classes m is even.

a. Find $\frac{n}{2}$

b. Identify the class whose cumulative frequency contains this $\frac{n}{2}$, and take the next class as the median class.

Illustration 6.7: Table 6.5 shows the grouped frequency distribution of the life of 350 tyres in 1000 km. Compute the median of the life of tyres.

Table 6.5 frequency distribution life of 350 tyres

	Given		*Computed*
Cell Number m	*Cell Interval*	*frequency f_i*	$\sum fi$
1	23.5 – 26.5	4	4
2	26.5 – 29.5	36	40
3	29.5 – 32.5	51	91
4	32.5 – 35.5	63	154
5	35.5 – 38.5	58	212
6	38.5 – 41.5	52	264
7	41.5 – 44.5	34	298
8	44.5 – 47.5	23	321
9	47.5 – 50.5	19	340
10	50.5 – 53.5	10	350
Total		$n = 350$	

Solution:

It can be observed from the frequency distribution Table 2.5 that the number of cells m is 10 which is an even number.

To Identify the median class

Step 1: $\frac{n}{2} = \frac{350}{2} = 175$

The value 175 is found in the cumulative frequency of 5th class and next class i.e., 6th class with class internal (38.5–41.5) is taken as the median class

Here $L_m = 38.5$; $f_m = 58$; $Cf_m = 264$; $i = 3$

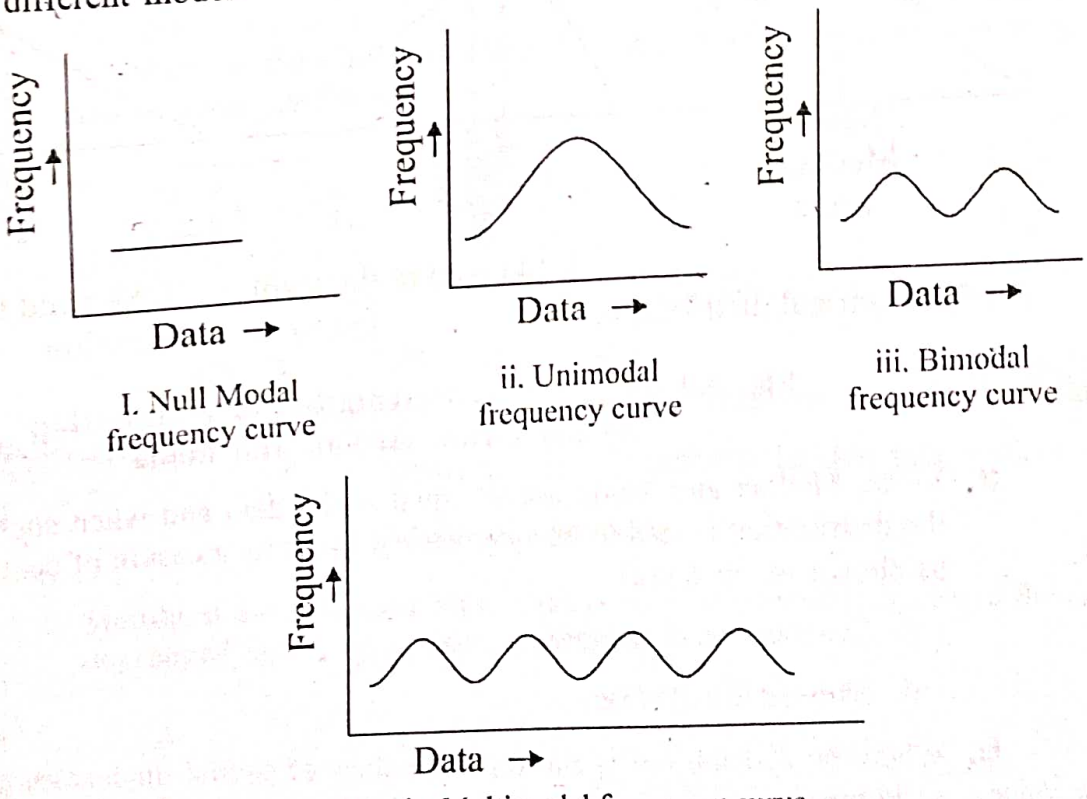
Then, $M_d = L_m + i\left(\frac{\frac{n}{2} - cf_m}{f_m}\right)$

$$= 38.5 + 3\left(\frac{\frac{350}{2} - 264}{58}\right) = 33.896 \simeq 33.9 \ (in\ 1000\ km)$$

c. **Mode:** Mode of a set of number or data is that value of the data which exists with highest frequency. A set of data may not have mode at all i.e.; all the values in the data exists with the same frequency.

Illustration 6.8: The set of numbers 2,3,7, 9,20 or 2,2, 3,3, 7,7, 9,9, 20,20 or 2,2,2 3,3,3, 7,7,7, 9,9,9, 20,20,20 have all the numbers with the same frequency. Such a set of data is said to be null modal data. A data may have only one mode.

For instance, the set of data such as 28,29,28,29,30,28,41,42,30,41 has the numbers 28 occurring three times which is highest such a set of data are called *unimodal* data. If there are more than two modes in a set of data than such a data is called *bimodal* data and more than two modes are called multimodal data. Fig 6.8 shows the different types of frequency curves with reference to different modes.



I. Null Modal frequency curve

ii. Unimodal frequency curve

iii. Bimodal frequency curve

iv. Multimodal frequency curve

Fig. 6.8 Frequency curves with different modes

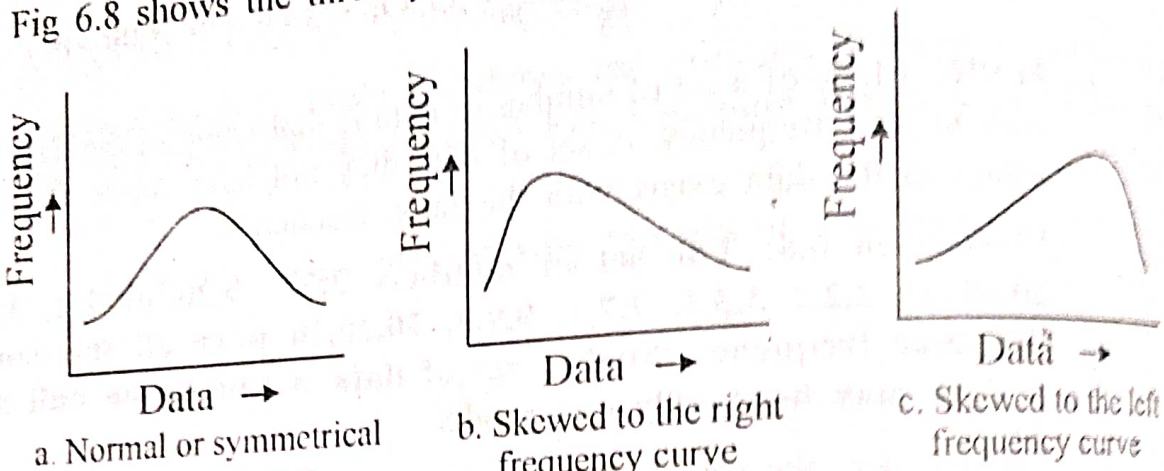
Other types of frequency curves are

a. Symmetrical or Normal curve

b. Skewed to the right

c. Skewed to the left

Fig 6.8 shows the three types of frequency curves



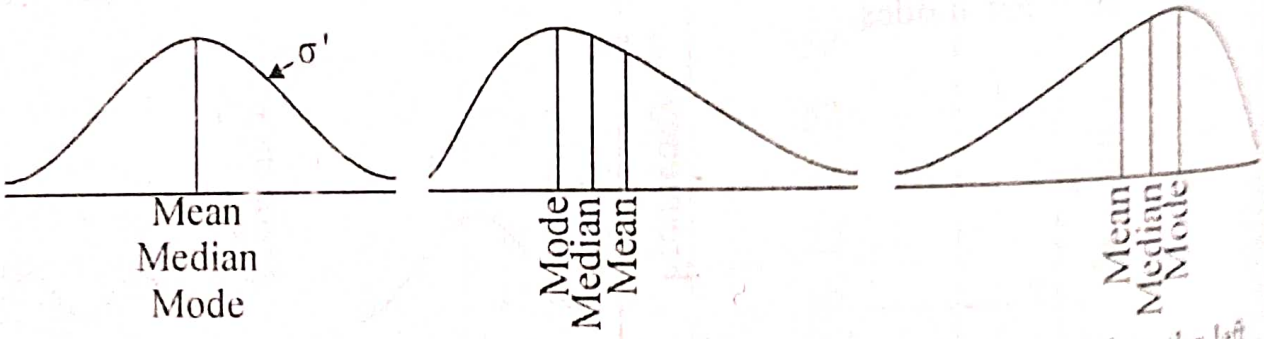
a. Normal or symmetrical frequency curve

b. Skewed to the right frequency curve

c. Skewed to the left frequency curve

Fig. 6.8 Types of Frequency curves

Fig 6.9 shows the relationship among the three methods of measure of central tendency with the help of frequency curves. The two possibilities are



i. Skewed to the right ii. Skewed to the left

(a) Symmetrical distribution (b) Skew distribution

Fig. 6.9 Graphical representation of relationship among mean, median and mode

a. Mean, Median and Mode are all equal in the data and when such is the case the distribution is said to be symmetrical about its measure of central tendency as shown in fig 6.9(a)

i. Symmetrical distribution

ii. Skewed distribution

b. When the distribution is skewed the values of central tendencies are different as shown in fig 6.9(b)

Note: *Mean is the most commonly used measure of central tendency*

MEASURE OF DISPERSION

The other numerical quantity used to describe a date is *measure of dispersion.* Measure of dispersion (spread) indicates how a data is spread about or scattered measure of central tendency. As mentioned earlier both measure of dispersion and measured of dispersion are required to describe a set of data.

The most commonly used methods to measure the dispersion are

a. Range b. Standard deviation

a. Range: The range of a set of data is the difference between the highest and lowest values in the data, and is denoted by R. R is given by the formula

$R = (X_H - X_L)$ Where, R = Range

X_H = Highest value in the data

X_L = Lowest value in the data

Illustration 6.9: The height in mts of five students in a class are as follows 1.79, 1.62, 1.83, 1.91, 1.57

Here the $X_L = 1.57$ $X_H = 1.91$

Range R = (1.91 – 1.57) = **0.34 mts**

Range is the simplest and easier to calculate.

b. Standard Deviation: The most commonly used measure of dispersion is *Standard Deviation.* Standard deviation is the root mean square (RMS) deviation of the observed value. Standard deviation is denoted by σ (sigma) and is given by the formula

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}$$

where, σ = standard deviation

n = total number of observation in the set of data

X_i = value of the ith number in the data

$\overline{X}$ = Mean of the data

Standard deviation for Raw data (Ungrouped data): Standard deviation for ungrouped data is given by the formula

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}} \text{ or } \sigma = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2} = \sqrt{\text{Mean of square} - \text{Square of mean}}$$

Illustration 6.10: The height of Six Students in a class is as shown in the table 6.6. Find the Standard deviation of the height of the students.

Table 6.6 Height of Six Students in mts

	Given	Computed
Sl.No.	*height*	X_i^2
1	1.7	2.89
2	1.6	2.56
3	1.5	2.25
4	1.8	3.24
5	2.0	4.00
6	1.4	1.96
$n=6$	$\sum X_i = 10$	$\sum X_i^2 = 16.9$

Solution:

$$Mean = \overline{X} = \left(\frac{10}{6}\right) = 1.67$$

$$Mean\ of\ X_i^2 = \left(\frac{\sum X_i^2}{n}\right) = \left(\frac{16.9}{6}\right) = 2.82$$

We have

$$\sigma = \sqrt{\frac{\left(\sum X_i^2\right)}{n} - \overline{X}^2} = \sqrt{2.82 - (1.67)^2} = \sqrt{2.82 - 2.56} = 0.26$$

Standard deviation of grouped data (using cell midpoint): Standard deviation of grouped data is given by the formula

$$\sigma = \sqrt{\frac{\sum_{j=1}^{m} f_j X_j^2}{n} - \frac{\left(\sum_{j=1}^{m} f_j X_j\right)^2}{n}}$$

where f_j = *frequency of the jth class*

m = *number of classes or cells*

X_j = *class midpoint of the jth class*

n = *total number of observations in the data*

Illustration 6.11: Table 6.7 shows the frequency distribution of electrical resistance of 500 fuses in ohms. Compute the standard deviation of the electrical resistance using cell midpoint.

Table 6.7 Frequency distribution of resistance of 500 fuses

Given			Computed			
Sl.No.	*Resistance Ohms*	*frequency* f_i	*Class mark* X_j	X_j^2	$f_j X_j^2$	$f_j X_j$
1	2.5 – 5.5	2	4	16	32	8
2	5.5 – 8.5	16	7	49	784	112
3	8.5 – 11.5	46	10	100	4600	460
4	11.5 – 14.5	88	13	169	14872	1144
5	14.5 – 17.5	138	16	256	35328	2208
6	17.5 – 20.5	113	19	361	40793	2147
7	20.5 – 23.5	71	22	484	34364	1562
8	23.5 – 26.5	22	25	625	13750	550
9	26.5 – 29.5	4	28	784	3136	112
		$\sum f_i = n = 500$			$\sum f_j x_j^2 = 16887$	$\sum f_j x_j = 8303$

We have for grouped data frequency distribution

$$\sigma = \sqrt{\left(\frac{\sum f_j x_j^2}{n}\right) - \left(\frac{\sum f_j x_j}{n}\right)^2} = \sqrt{\left(\frac{16887}{500}\right) - \left(\frac{8303}{500}\right)} = \sqrt{33.774 - 16.606}$$

$$= \sqrt{17.168} = \boxed{4.143}$$

Standard deviation of grouped data by deviation method: The standard deviation of grouped data is also given by the formula

$$\sigma = \left[\sqrt{\frac{\sum f_j d_j^2}{n} - \left(\frac{\sum f_j d_j}{n}\right)^2}\right] i$$

where, f_j = frequency of the jth class

i = class width

d_j = deviation by class of the jth class from the assumed mean class

n = total number of observations

Illustration 6.12: The data shown in table 6.8 is solved here by deviation method

Table : 6.8

Given			Computed		
Class no.	Class interval Resistance in ohms	frequency f_i	Deviation d_i	$f_i d_i$	$f_i d_i^2$
1	2.5 – 5.5	2	–4	–8	32
2	5.5 – 8.5	16	–3	–48	144
3	8.5 – 11.5	46	–2	–92	184
4	11.5 – 14.5	88	–1	–88 (–236)	88
5	14.5 – 17.5	138	0	0	0
6	17.5 – 20.5	113	1	113	113
7	20.5 – 23.5	71	2	142	284
8	23.5 – 26.5	22	3	66	198
9	26.5 – 29.5	4	4	16 (+337)	64
		$\sum f_i = 500$		$\sum f_i d_i = +101$	$\sum f_i d_i^2 = 1107$

Solution:

Let us assume 5th class as the median class and write the deviation of all other classes.

Then $\sum f_i d_i$ and $\sum f_i d_i^2$ are computed. Table 2.12 shows the computed values for each class and the corresponding $\sum f_i d_i$ and $\sum f_i d_i^2$.

Class width $i = (5.5 - 2.5) = 3$

Then standard deviation is computed using the formula

$$\sigma = i\left[\sqrt{\frac{\sum f_i d_i^2}{n} - \left(\frac{f_i d_i}{n}\right)^2}\right]$$

$$= 3\left[\sqrt{\frac{1107}{500} - \left(\frac{101}{500}\right)^2}\right] = 3\left[\sqrt{2.214 - 0.04}\right]$$

$$= 3\left(\sqrt{2.174}\right) = \mathbf{4.42\ ohms}$$