CHAPTER 14 Statistics

14.1 Introduction

Ganesh had recorded the marks of 26 children in his class in the mathematics Summative Assessment - I in the register as follows:

Arjun	76	Narayana	12
Kamini	82	Suresh	24
Shafik	64	Durga	39
Keshav	53	Shiva	41
Lata	90	Raheem	69
Rajender	27	Radha	73
Ramu	34	Kartik	94
Sudha	74	Joseph	89
Krishna	76	Ikram	64
Somu	65	Laxmi	46
Gouri	47	Sita	19
Upendra	54	Rehana	53
Ramaiah	36	Anitha	69

Is the data given in the list above organized? Why or why not?

His teacher asked him to report on how his class students had performed in mathematics in their Summative Assessment - I .

Ganesh made the following table to understand the performance of his class:

Marks	Number of children
0 - 33	4
34 - 50	6
51 - 75	10
76 - 100	6

Is the data given in the above table grouped or ungrouped?

He showed this table to his teacher and the teacher appreciated him for organising the data to be understood easily. We can see that most children have got marks between 51-75. Do you think Ganesh should have used smaller range? Why or why not?

In the previous class, you had learnt about the difference between grouped and ungrouped data as well as how to present this data in the form of tables. You had also learnt to calculate the mean value for ungrouped data. Let us recall this learning and then learn to calculate the mean, median and mode for grouped data.

14.2 Mean of Ungrouped Data

As we know the mean (or average) of observations is the sum of the values of all the observations divided by the total number of observations. Let $x_1, x_2, \ldots, x_n$ be observations with respective frequencies $f_1, f_2, \ldots, f_n$. This means that observation x_1 occurs f_1 times, x_2 occurs f_2 times, and so on.

Now, the sum of the values of all the observations $= f_1x_1 + f_2x_2 + \ldots + f_nx_n$, and the number of observations $= f_1 + f_2 + \ldots + f_n$.

So, the mean $\bar{x}$ of the data is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots\ldots + f_nx_n}{f_1 + f_2 + \ldots\ldots\ldots + f_n}$$

Recall that we can write this in short, using the Greek letter $\sum$ which means summation

i.e., $\bar{x} = \frac{\sum f_ix_i}{\sum f_i}$

Example-1. The marks obtained in mathematics by 30 students of Class X of a certain school are given in table below. Find the mean of the marks obtained by the students.

Marks obtained (x_i)	10	20	36	40	50	56	60	70	72	80	88	92	95
Number of student (f_i)	1	1	3	4	3	2	4	4	1	1	2	3	1

Solution : Let us re-organize this data and find the sum of all observations.

Marks obtained (x_i)	Number of students (fi)	$f_i x_i$
10	1	10
20	1	20
36	3	108
40	4	160
50	3	150
56	2	112
60	4	240
70	4	280
72	1	72
80	1	80
88	2	176
92	3	276
95	1	95
Total	$\sum f_i$ =30	$\sum f_i x_i$ = 1779

So, $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1779}{30} = 59.3$

Therefore, the mean marks are 59.3.

In most of our real life situations, data is usually so large that to make a meaningful study, it needs to be condensed as grouped data. So, we need to convert ungrouped data into grouped data and devise some method to find its mean.

Let us convert the ungrouped data of Example 1 into grouped data by forming class-intervals of width, say 15. Remember that while allocating frequencies to each class-interval, students whose score is equal to in any **upper class-boundary** would be considered in the next class, e.g., 4 students who have obtained 40 marks would be considered in the class-interval 40-55 and not in 25-40. With this convention in our mind, let us form a grouped frequency distribution table.

Class interval	10-25	25-40	40-55	55-70	70-85	85-100
Number of students	2	3	7	6	6	6

Now, for each class-interval, we require a point which would serve as the representative of the whole class. ***It is assumed that the frequency of each class-interval is centred around its mid-point.*** So, the *mid-point* of each class can be chosen to represent the observations falling in that class and is called the class mark. Recall that we find the class mark by finding the average of the upper and lower limit of the class.

$$\text{Class mark} = \frac{\text{Upper class limit} + \text{Lower class limit}}{2}$$

For the class 10 -25, the class mark is $\frac{10+25}{2}$ =17.5. Similarly, we can find the class marks of the remaining class intervals. We put them in the table. These class marks serve as our x_i's. We can now proceed to compute the mean in the same manner as in the previous example.

Class interval	Number of students (f_i)	Class Marks (x_i)	$f_i x_i$
10-25	2	17.5	35.0
25-40	3	32.5	97.5
40-55	7	47.5	332.5
55-70	6	62.5	375.0
70-85	6	77.5	465.0
85-100	6	92.5	555.0
Total	$\sum f_i$ =30		$\sum f_i x_i$ =1860.0

The sum of the values in the last column gives us $\sum f_i x_i$. So, the mean $\bar{x}$ of the given data is given by

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1860}{30} = 62$$

This new method of finding the mean is known as the **Direct Method**.

We observe that in the above cases we are using the same data and employing the same formula for calculating the mean but the results obtained are different. In example (1), 59.3 is the exact mean and 62 is the approximate mean. Can you think why this is so?

THINK - DISCUSS

1. The mean value can be calculated from both ungrouped and grouped data. Which one do you think is more accurate? Why?
2. When it is more convenient to use grouped data for analysis?

Sometimes when the numerical values of x_1 and f_1 are large, finding the product of x_1 and f_1 becomes tedious and time consuming. So, for such situations, let us think of a method of reducing these calculations.

We can do nothing with the f_i's, but we can change each x_i to a smaller number so that our calculations become easy. How do we do this? What is about subtracting a fixed number from each of these x_i's? Let us try this method for the data in example 1.

The first step is to choose one among the x_i's as the *assumed mean*, and denote it by 'a'. Also, to further reduce our calculation work, we may take 'a' to be that x_i which lies in the centre of $x_1, x_2, ..., x_n$. So, we can choose $a = 47.5$ or $a = 62.5$. Let us choose $a = 47.5$.

The second step is to find the **deviation** of 'a' from each of the x_i's, which we denote as d_i

i.e., $d_i = x_i - a = x_i - 47.5$

The third step is to find the product of d_i with the corresponding f_i, and take the sum of all the $f_i d_i$'s. These calculations are shown in table given below-

Class interval	Number of students (f_i)	Class Marks (x_i)	$d_i = x_i - 47.5$ $x_i = a$	$f_i d_i$
10-25	2	17.5	-30	-60
25-40	3	32.5	-15	-45
40-55	7	47.5 (a)	0	0
55-70	6	62.5	15	90
70-85	6	77.5	30	180
85-100	6	92.5	45	270
Total	$\sum f_i = 30$			$\sum f_i d_i = 435$

So, from the above table, the mean of the deviations, $\bar{d} = \frac{\sum f_i d_i}{\sum f_i}$

Now, let us find the relation between $\bar{d}$ and $\bar{x}$.

Since, in obtaining d_i we subtracted 'a' from each x_i so, in order to get the mean $\bar{x}$ we need to add 'a' to $\bar{d}$. This can be explained mathematically as:

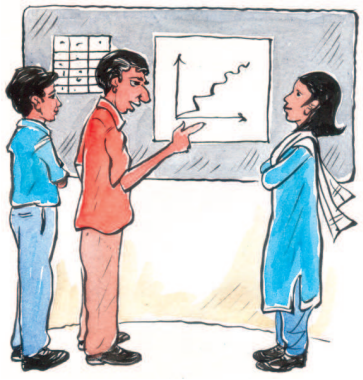
Mean of deviations, $\bar{d} = \frac{\sum f_i d_i}{\sum f_i}$

So, $\bar{d} = \frac{\sum f_i (x_i - a)}{\sum f_i}$

$$= \frac{\sum f_i x_i}{\sum f_i} - \frac{\sum f_i a}{\sum f_i}$$

$$= \bar{x} - a\frac{\sum f_i}{\sum f_i}$$

$$\bar{d} = \bar{x} - a$$



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

Substituting the values of a, $\sum f_i d_i$ and $\sum f_i$ from the table, we get

$$\bar{x} = 47.5 + \frac{435}{30} = 47.5 + 14.5 = 62$$

Therefore, the mean of the marks obtained by the students is 62.

The method discussed above is called the **Assumed Mean Method**.

ACTIVITY

Consider the data given in example 1 and calculate the arithmetic mean by deviation method by taking successive values of x_i i.e., 17.5, 32.5, ... as assumed means. Now discuss the following:

1. Are the values of arithmetic mean in all the above cases equal?
2. If we take the actual mean as the assumed mean, how much will $\sum f_i d_i$ be?
3. Reason about taking any mid-value (class mark) as assumed mean?

Observe that in the table given below the values in Column 4 are all multiples of 15. So, if we divide all the values of Column 4 by 15, we would get smaller numbers which we then multiply with f_i. (Here, 15 is the class size of each class interval.)

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

Now, we calculate u_i in this way and continue as before (i. e., find $f_i u_i$ and then $\sum f_i u_i$). Taking $h = 15$, [Generally size of the class is taken as h but it need not be size of the class always].

Let $\bar{u} = \dfrac{\sum f_i u_i}{\sum f_i}$

Class interval	Number of students (f_i)	Class Marks (x_i)	$d_i = x_i - a$	$u_i = \dfrac{x_i - a}{h}$	$f_i u_i$
10-25	2	17.5	-30	-2	-4
25-40	3	32.5	-15	-1	-3
40-55	7	47.5	0	0	0
55-70	6	62.5	15	1	6
70-85	6	77.5	30	2	12
85-100	6	92.5	45	3	18
Total	$\sum f_i$ =30				$\sum f_i u_i$ =29

Here, again let us find the relation between $\bar{u}$ and $\bar{x}$.

We have $u_i = \dfrac{x_i - a}{h}$

So $\bar{u} = \dfrac{\sum f_i u_i}{\sum f_i}$

So $\bar{u} = \dfrac{\sum f_i \dfrac{(x_i - a)}{h}}{\sum f_i}$

$$= \frac{1}{h}\left\{\frac{\sum f_i x_i}{\sum f_i} - \frac{\sum f_i a}{\sum f_i}\right\}$$

$$= \frac{1}{h}(\bar{x} - a)$$

or $h\bar{u} = \bar{x} - a$

$$\bar{x} = a + h\bar{u}$$

Therefore, $\bar{x} = a + h\left\{\dfrac{\sum f_i u_i}{\sum f_i}\right\}$

or $$\bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \times h$$

Substituting the values of a, $\sum f_i u_i$ and $\sum f_i$ from the table, we get

$$\bar{x} = 47.5 + 15 \times \frac{29}{30}$$

$$= 47.5 + 14.5 = 62$$

So, the mean marks obtained by a student are 62.

The method discussed above is called the **Step-deviation** method.

We note that:

- The step-deviation method will be convenient to apply if all the d_i's have a common factor.
- The mean obtained by all the three methods is the same.
- The assumed mean method and step-deviation method are just simplified forms of the direct method.
- The formula $\bar{x} = a + h\bar{u}$ still holds if a and h are not as given above, but are any non-zero numbers such that $u_i = \frac{x_i - a}{h}$

Let us apply these methods in other examples.

Example-2. The table below gives the percentage distribution of female teachers in the primary schools of rural areas of various states and union territories (U.T.) of India. Find the mean percentage of female teachers using all the three methods.

Percentage of female teachers	15 - 25	25 - 35	35 – 45	45 - 55	55 - 65	65 - 75	75 – 85
Number of States/U.T.	6	11	7	4	4	2	1

Source : *Seventh All India School Education Survey conducted by NCERT*

Solution : Let us find the class marks x_i of each class, and put them in a table

Here we take $a = 50$, $h = 10$,

then $d_i = x_i - 50$ and $u_i = \frac{x_i - 50}{10}$

We now find d_i and u_i and put them in the table

Percentage of female teachers	Number of States/U.T.	x_i	$d_i = x_i - 50$	$u_i = \frac{x_i - 50}{10}$	$f_i x_i$	$f_i d_i$	$f_i u_i$
15 – 25	6	20	-30	-3	120	-180	-18
25 – 35	11	30	-20	-2	330	-220	-22
35 – 45	7	40	-10	-1	280	-70	-7
45 – 55	4	50	0	0	200	0	0
55 – 65	4	60	10	1	240	40	4
65 – 75	2	70	20	2	140	40	4
75 – 85	1	80	30	3	80	30	3
Total	35				1390	-360	-36

From the table above, we obtain $\sum f_i = 35$, $\sum f_i x_i = 1390$, $\sum f_i d_i = -360$, $\sum f_i u_i = -36$.

Using the direct method $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1390}{35} = 39.71$

Using the assumed mean method $\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i} = 50 + \frac{-360}{35} = 50 - 10.29 = 39.71$

Using the step-deviation method $\bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \times h = 50 + \frac{-36}{35} \times 10 = 39.71$

Therefore, the mean percentage of female teachers in the primary schools of rural areas is 39.71.

THINK - DISCUSS

1. Is the result obtained by all the three methods the same?
2. If x_i and f_i are sufficiently small, then which method is an appropriate choice?
3. If x_i and f_i are numerically large numbers, then which methods are appropriate to use?

Even if the class sizes are unequal, and x_i are large numerically, we can still apply the step-deviation method by taking h to be a suitable divisor of all the d_i's.

Example-3. The distribution below shows the number of wickets taken by bowlers in one-day cricket matches. Find the mean number of wickets by choosing a suitable method. What does the mean signify?

Number of wickets	20 - 60	60 - 100	100 - 150	150 - 250	250 - 350	350 – 450
Number of bowlers	7	5	16	12	2	3

Solution : Here, the class size varies, and the x_i's are large. Let us still apply the step deviation method with $a = 200$ and $h = 20$. Then, we obtain the data as given in the table.

Number of wickets	Number of bowlers (f_i)	x_i	$d_i = x_i - a$	$u_i = \frac{x_i - a}{h}$ ($h = 20$)	$f_i u_i$
20 – 60	7	40	-160	-8	-56
60 – 100	5	80	-120	-6	-30
100 – 150	16	125	-75	-3.75	-60
150 – 250	12	200 (a)	0	0	0
250 – 350	2	300	100	5	10
350 – 450	3	400	200	10	30
Total	45				-106

So $\bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \times h = 200 + \frac{-106}{45} \times 20 = 200 - 47.11 = 152.89$

Thus, the average number of wickets taken by these 45 bowlers in one-day cricket is 152.89.

Classroom Project :

1. Collect the marks obtained by all the students of your class in Mathematics in the recent examination conducted in your school. Form a grouped frequency distribution of the data obtained. Do the same regarding other subjects and compare. Find the mean in each case using a method you find appropriate.
2. Collect the daily maximum temperatures recorded for a period of 30 days in your city. Present this data as a grouped frequency table. Find the mean of the data using an appropriate method.
3. Measure the heights of all the students of your class and form a grouped frequency distribution table of this data. Find the mean of the data using an appropriate method.

EXERCISE - 14.1

1. A survey was conducted by a group of students as a part of their environment awareness programme, in which they collected the following data regarding the number of plants in 20 houses in a locality. Find the mean number of plants per house.

Number of plants	0 - 2	2 - 4	4 – 6	6 - 8	8 - 10	10 - 12	12 – 14
Number of houses	1	2	1	5	6	2	3

2. Consider the following distribution of daily wages of 50 workers of a factory.

Daily wages in Rupees	200 - 250	250 - 300	300 - 350	350 - 400	400– 450
Number of workers	12	14	8	6	10

Find the mean daily wages of the workers of the factory by using an appropriate method.

3. The following distribution shows the daily pocket allowance of children of a locality. The mean pocket allowance is ₹ 18. Find the missing frequency f.

Daily pocket allowance(in Rupees)	11 - 13	13 - 15	15 - 17	17 - 19	19 - 21	21 - 23	23 - 25
Number of children	7	6	9	13	f	5	4

4. Thirty women were examined in a hospital by a doctor and their of heart beats per minute were recorded and summarised as shown. Find the mean heart beats per minute for these women, choosing a suitable method.

Number of heart beats/minute	65-68	68-71	71-74	74-77	77-80	80-83	83-86
Number of women	2	4	3	8	7	4	2

5. In a retail market, fruit vendors were selling oranges kept in packing baskets. These baskets contained varying number of oranges. The following was the distribution of oranges.

Number of oranges	10-14	15–19	20-24	25-29	30–34
Number of baskets	15	110	135	115	25

Find the mean number of oranges kept in each basket. Which method of finding the mean did you choose?

6. The table below shows the daily expenditure on food of 25 households in a locality.

Daily expenditure (in Rupees)	100-150	150-200	200-250	250-300	300-350
Number of house holds	4	5	12	2	2

Find the mean daily expenditure on food by a suitable method.

7. To find out the concentration of SO_2 in the air (in parts per million, i.e., ppm), the data was collected for 30 localities in a certain city and is presented below:

Concentration of SO_2 in ppm	0.00-0.04	0.04-0.08	0.08-0.12	0.12-0.16	0.16-0.20	0.20-0.24
Frequency	4	9	9	2	4	2

Find the mean concentration of SO_2 in the air.

8. A class teacher has the following attendance record of 40 students of a class for the whole term. Find the mean number of days a student was present out of 56 days in the term.

Number of days	35-38	38-41	41-44	44-47	47-50	50-53	53-56
Number of students	1	3	4	4	7	10	11

9. The following table gives the literacy rate (in percentage) of 35 cities. Find the mean literacy rate.

Literacy rate in %	45–55	55-65	65-75	75-85	85-95
Number of cities	3	10	11	8	3

14.3 Mode

A mode is that value among the observations which occurs most frequently.

Before learning about calculating the mode of grouped data let us first recall how we found the mode for ungrouped data through the following example.

Example-4. The wickets taken by a bowler in 10 cricket matches are as follows: 2, 6, 4, 5, 0, 2, 1, 3, 2, 3. Find the mode of the data.

Solution : Let us arrange the observations in order i.e., 0, 1, 2, 2, 2, 3, 3, 4, 5, 6

Clearly, 2 is the number of wickets taken by the bowler in the maximum number of matches (i.e., 3 times). So, the mode of this data is 2.

Do This

1. Find the mode of the following data.
   a) 5, 6, 9, 10, 6, 12, 3, 6, 11, 10, 4, 6, 7.
   b) 20, 3, 7, 13, 3, 4, 6, 7, 19, 15, 7, 18, 3.
   c) 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6.
2. Is the mode always at the centre of the data?
3. Does the mode change. If another observation is added to the data in Example? Comment.
4. If the maximum value of an observation in the data in Example 4 is changed to 8, would the mode of the data be affected? Comment.

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

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

where, l = lower boundary of the modal class,

h = size of the modal class interval,

f_1 = frequency of the modal class,

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class.

Let us consider the following examples to illustrate the use of this formula.

Example-5. A survey conducted on 20 households in a locality by a group of students resulted in the following frequency table for the number of family members in a household.

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

Find the mode of this data.

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

Now,

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

frequency of the modal class (f_1) = 8,

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

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

Now, let us substitute these values in the formula-

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

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

Therefore, the mode of the data above is 3.286.

Example-6. The marks distribution of 30 students in a mathematics examination are given in the adjacent table. Find the mode of this data. Also compare and interpret the mode and the mean.

Class interval	Number of students (f_i)	Class Marks (x_i)	$f_i x_i$
10-25	2	17.5	35.0
25-40	3	32.5	97.5
40-55	7	47.5	332.5
55-70	6	62.5	375.0
70-85	6	77.5	465.0
85-100	6	92.5	555.0
Total	$\sum f_i = 30$		$\sum f_i x_i = 1860.0$

Solution : Since the maximum number of students (i.e., 7) have got marks in the interval, 40-65 the modal class is 40 - 55.

The lower boundary (l) of the modal class = 40,

The class size (h) = 15,

The frequency of modal class (f_1) = 7,

the frequency of the class preceding the modal class (f_0) = 3,

the frequency of the class succeeding the modal class (f_2) = 6.

Now, using the formula:

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

$$= 40 + \left(\frac{7-3}{2\times 7 - 6 - 3}\right) \times 15 = 40 + 12 = 52$$

Interpretation : The mode marks is 52. Now, from Example 1, we know that the mean marks is 62. So, the maximum number of students obtained 52 marks, while on an average a student obtained 62 marks.

THINK - DISCUSS

1. It depends upon the demand of the situation whether we are interested in finding the average marks obtained by the students or the marks obtained by most of the students.
   a. What do we find in the first situation?
   b. What do we find in the second situation?
2. Can mode be calculated for grouped data with unequal class sizes?

EXERCISE - 14.2

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

Age (in years)	5-15	15-25	25-35	35-45	45-55	55-65
Number of patients	6	11	21	23	14	5

Find the mode and the mean of the data given above. Compare and interpret the two measures of central tendency.

2. The following data gives the information on the observed life times (in hours) of 225 electrical components :

Lifetimes (in hours)	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100	100 - 120
Frequency	10	35	52	61	38	29

Determine the modal lifetimes of the components.

3. The following data gives the distribution of total monthly household expenditure of 200 families of a village. Find the modal monthly expenditure of the families. Also, find the mean monthly expenditure :

Expenditure (in rupees)	1000-1500	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500	4500-5000
Number of families	24	40	33	28	30	22	16	7

4. The following distribution gives the state-wise, teacher-student ratio in higher secondary schools of India. Find the mode and mean of this data. Interpret the two measures.

Number of students	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55
Number of States	3	8	9	10	3	0	0	2

5. The given distribution shows the number of runs scored by some top batsmen of the world in one-day international cricket matches.

Runs	3000-4000	4000-5000	5000-6000	6000-7000	7000-8000	8000-9000	9000-10000	10000-11000
Number of batsmen	4	18	9	7	6	3	1	1

Find the mode of the data.

6. A student noted the number of cars passing through a spot on a road for 100 periods, each of 3 minutes, and summarised this in the table given below.

Number of cars	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80
Frequency	7	14	13	12	20	11	15	8

Find the mode of the data.

14.4 Median of Grouped Data

Median is a measure of central tendency which gives the value of the middle-most observation in the data. Recall that for finding the median of ungrouped data, we first arrange the data values or the observations in ascending order.

Then, if n is odd, the median is the $\left(\frac{n+1}{2}\right)^{th}$ observation and

if n is even, then the median will be the average of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations.

Suppose, we have to find the median of the following data, which is about the marks, out of 50 obtained by 100 students in a test :

Marks obtained	20	29	28	33	42	38	43	25
Number of students	6	28	24	15	2	4	1	20

First, we arrange the marks in ascending order and prepare a frequency table 14.9 as follows :

Marks obtained	Number of students (frequency)
20	6
25	20
28	24
29	28
33	15
38	4
42	2
43	1
Total	100

Here $n = 100$, which is even. The median will be the average of the $\left(\frac{n}{2}\right)^{th}$ and the $\left(\frac{n}{2}+1\right)^{th}$ observations, i.e., the 50^{th} and 51^{st} observations. To find the position of these middle values, we construct cumulative frequency.

Marks obtained	Number of students	Cumulative frequency
20	6	6
upto 25	6 + 20 = 26	26
upto 28	26 + 24 = 50	50
upto 29	50 + 28 = 78	78
upto 33	78 + 15 = 93	93
upto 38	93 + 4 = 97	97
upto 42	97 + 2 = 99	99
upto 43	99 + 1 = 100	100

Now we add another column depicting this information to the frequency table above and name it as *cumulative frequency column*.

From the table above, we see that :

50^{th} observation is 28 (Why?)

51^{st} observation is 29

$$\text{Median} = \frac{28+29}{2} = 28.5$$

Remark : Column 1 and column 3 in the above table are known as *Cumulative Frequency Table*. The median marks 28.5 conveys the information that about 50% students obtained marks less than 28.5 and another 50% students obtained marks more than 28.5.

Consider a grouped frequency distribution of marks obtained, out of 100, by 53 students, in a certain examination, as shown in adjacent table.

Marks	Number of students
0-10	5
10-20	3
20-30	4
30-40	3
40-50	3
50-60	4
60-70	7
70-80	9
80-90	7
90-100	8

From the table, try to answer the following questions :

How many students have scored marks less than 10? The answer is clearly 5.

How many students have scored less than 20 marks? Observe that the number of students who have scored less than 20 include the number of students who have scored marks from 0-10 as well as the number of students who have scored marks from 10-20. So, the total number of students with marks less than 20 is 5 + 3, i.e., 8. We say that the cumulative frequency of the class 10-20 is 8. (As shown in table 14.11)

Similarly, we can compute the cumulative frequencies of the other classes, i.e., the number of students with marks less than 30, less than 40, ..., less than 100.

Marks obtained	Number of students (Cumulative frequency)
Less than 10	5
Less than 20	5 + 3 = 8
Less than 30	8 + 4 = 12
Less than 40	12 + 3 = 15
Less than 50	15 + 3 = 18
Less than 60	18 + 4 = 22
Less than 70	22 + 7 = 29
Less than 80	29 + 9 = 38
Less than 90	38 + 7 = 45
Less than 100	45 + 8 = 53

This distribution is called the cumulative frequency distribution of the less than type. Here 10, 20, 30, ..., 100, are the upper boundaries of the respective class intervals.

We can similarly make the table for the number of students with scores more than or equal to 0 (this number is same as sum of all the frequencies), more than above sum minus the frequency of the first class interval), more than or equal to 20 (this number is same as the sum of all frequencies minus the sum of the frequencies of the first two class intervals), and so on. We observe that all 53 students have scored marks more than or equal to 0. Since there are 5 students scoring marks in the interval 0-10, this means that there

Marks obtained	Number of students (Cumulative frequency)
More than or equal to 0	53
More than or equal to 10	53 - 5 = 48
More than or equal to 20	48 - 3 = 45
More than or equal to 30	45 - 4 = 41
More than or equal to 40	41 - 3 = 38
More than or equal to 50	38 - 3 = 35
More than or equal to 60	35 - 4 = 31
More than or equal to 70	31 - 7 = 24
More than or equal to 80	24 - 9 = 15
More than or equal to 90	15 - 7 = 8

are 53-5 = 48 students getting more than or equal to 10 marks. Continuing in the same manner, we get the number of students scoring 20 or above as 48-3 = 45, 30 or above as 45-4 = 41, and so on, as shown in the table a side.

This table above is called a cumulative frequency distribution of the more than type. Here 0, 10, 20, ..., 90 give the lower boundaries of the respective class intervals.

Now, to find the median of grouped data, we can make use of any of these cumulative frequency distributions.

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

To find this class, we find the cumulative frequencies of all the clases and $\frac{n}{2}$. We now locate the class whose cumulative frequency exceeds $\frac{n}{2}$ for the first time. This is called the median class.

Marks	Number of students (f)	Cumulative frequency (cf)
0-10	5	5
10-20	3	8
20-30	4	12
30-40	3	15
40-50	3	18
50-60	4	22
60-70	7	29
70-80	9	38
80-90	7	45
90-100	8	53

In the distribution above, n = 53. So $\frac{n}{2}$ = 26.5. Now 60-70 is the class whose cumulative frequency 29 is greater than (and nearest to) $\frac{n}{2}$, i.e., 26.5.

Therefore, 60-70 is the median class.

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

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

where l = lower boundary of median clas,

n = number of observations,

cf = cumulative frequency of class preceding the median class,

f = frequency of median class,

h = class size (size of the median class).

Substituting the values $\frac{n}{2} = 26.5$, $l = 60$, $cf = 22$, $f = 7$, $h = 10$

in the formula above, we get

$$\text{Median} = 60 + \left[\frac{26.5 - 22}{6} \right] \times 10$$

$$= 60 + \frac{45}{7}$$

$$= 66.4$$

So, about half the students have scored marks less than 66.4, and the other half have scored marks more than 66.4.

Example-7. A survey regarding the heights (in cm) of 51 girls of Class X of a school was conducted and data was obtained as shown in table. Find their median.

Height (in cm)	Number of girls
Less than 140	4
Less than 145	11
Less than 150	29
Less than 155	40
Less than 160	46
Less than 165	51

Solution : To calculate the median height, we need to find the class intervals and their corresponding frequencies. The given distribution being of the *less than type*, 140, 145, 150, . . ., 165 give the upper limits of the corresponding class intervals. So, the classes should be below 140, 140 - 145, 145 - 150, . . ., 160 - 165.

Class intervals	Frequency	Cumulative frequency
Below 140	4	4
140-145	7	11
145-150	18	29
150-155	11	40
155-160	6	46
160-165	5	51

Observe that from the given distribution, we find that there are 4 girls with height less than 140, i.e., the frequency of class interval below 140 is 4 . Now, there are 11 girls with heights less than 145 an d 4 girls with height less than 140. Therefore, the number of girls with height in the interval 140 - 145 is $11 - 4 = 7$. Similarly, the frequencies can be calculated as shown in table.

Number of observations, $n = 51$

$\frac{n}{2} = \frac{51}{2} = 25.5^{th}$ observation, which lies in the class 145 - 150.

$\therefore$ 145 - 150 is median class

Then, l (the lower boundary) = 145,

cf (the cumulative frequency of the class preceding 145 - 150) = 11,

f (the frequency of the median class 145 - 150) = 18,

h (the class size) = 5.

Using the formula, Median $= l + \frac{\left(\frac{n}{2} - cf\right)}{f} \times h$

$$= 145 + \frac{(25.5 - 11)}{18} \times 5$$

$$= 145 + \frac{72.5}{18} = 149.03$$

So, the median height of the girls is 149.03 cm. This means that the height of about 50% of the girls is less than this height, and that of other 50% is greater than this height.

Example-8. The median of the following data is 525. Find the values of x and y, if the total frequency is 100. Here, CI stands for class interval and Fr for frequency.

CI	0-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800	800-900	900-1000
Fr	2	5	x	12	17	20	y	9	7	4

Solution :

It is given that $n = 100$

So, $76 + x + y = 100$, i.e., $x + y = 24$ (1)

The median is 525, which lies in the class 500 – 600

So, $l = 500$, $f = 20$, $cf = 36 + x$, $h = 100$

Using the formula

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

$$525 = 500 + \frac{50 - 36 - x}{20} \times 100$$

i.e., $525 - 500 = (14 - x) \times 5$

i.e., $25 = 70 - 5x$

i.e., $5x = 70 - 25 = 45$

So, $x = 9$

Therefore, from (1), we get $9 + y = 24$

i.e., $y = 15$

Class intervals	Frequency	Cumulative frequency
0-100	2	2
100-200	5	7
200-300	x	$7+x$
300-400	12	$19+x$
400-500	17	$36+x$
500-600	20	$56+x$
600-700	y	$56+x+y$
700-800	9	$65+x+y$
800-900	7	$72+x+y$
900-1000	4	$76+x+y$

Note :

The median of grouped data with unequal class sizes can also be calculated.

14.5 Which value of Central Tendency

Which measure would be best suited for a particular requirement.

The mean is the most frequently used measure of central tendency because it takes into account all the observations, and lies between the extremes, i.e., the largest and the smallest observations of the entire data. It also enables us to compare two or more distributions. For example, by comparing the average (mean) results of students of different schools of a particular examination, we can conclude which school has a better performance.

However, extreme values in the data affect the mean. For example, the mean of classes having frequencies more or less the same is a good representative of the data. But, if one class has frequency, say 2, and the five others have frequency 20, 25, 20, 21, 18, then the mean will certainly not reflect the way the data behaves. So, in such cases, the mean is not a good representative of the data.

In problems where individual observations are not important, especially extreem values, and we wish to find out a 'typical' observation, the median is more appropriate, e.g., finding the typical productivity rate of workers, average wage in a country, etc. These are situations where extreme values may exist. So, rather than the mean, we take the median as a better measure of central tendency.

In situations which require establishing the most frequent value or most popular item, the mode is the best choice, e.g., to find the most popular T.V. programme being watched, the consumer item in greatest demand, the colour of the vehicle used by most of the people, etc.

EXERCISE - 14.3

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

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

2. If the median of 60 observations, given below is 28.5, find the values of x and y.

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

3. A life insurance agent found the following data about distribution of ages of 100 policy holders. Calculate the median age. [Policies are given only to persons having age 18 years onwards but less than 60 years.]

Age (in years)	Below 20	Below 25	Below 30	Below 35	Below 40	Below 45	Below 50	Below 55	Below 60
Number of policy holders	2	6	24	45	78	89	92	98	100

4. The lengths of 40 leaves of a plant are measured correct to the nearest millimetre, and the data obtained is represented in the following table :

Length (in mm)	118-126	127-135	136-144	145-153	154-162	163-171	172-180
Number of leaves	3	5	9	12	5	4	2

Find the median length of the leaves. (**Hint :** The data needs to be converted to continuous classes for finding the median, since the formula assumes continuous classes. The classes then change to 117.5 - 126.5, 126.5 - 135.5, . . ., 171.5 - 180.5.)

5. The following table gives the distribution of the life-time of 400 neon lamps

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

Find the median life time of a lamp.

6. 100 surnames were randomly picked up from a local telephone directory and the frequency distribution of the number of letters in the English alphabet in the surnames was obtained as follows

Number of letters	1-4	4-7	7-10	10-13	13-16	16-19
Number of surnames	6	30	40	16	4	4

Determine the median number of letters in the surnames. Find the mean number of letters in the surnames? Also, find the modal size of the surnames.

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

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

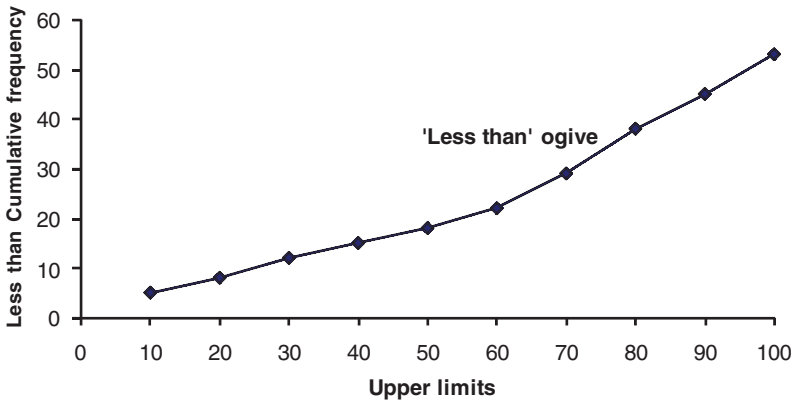
14.6 Graphical Representation of Cumulative Frequency Distribution

As we all know, pictures speak better than words. A graphical representation helps us in understanding given data at a glance. In Class IX, we have represented the data through bar graphs, histograms and frequency polygons. Let us now represent a cumulative frequency distribution graphically.

For example, let us consider the cumulative frequency distribution given in example.

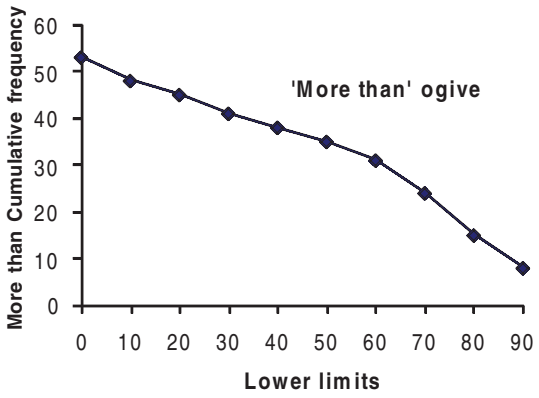
For drawing ozires, it should be ensured that the class intervals are continuous, because cumulative frequencies are linked with boundaries, but not with limits.

Recall that the values 10, 20, 30, ..., 100 are the upper boundaries of the respective class intervals. To represent the data graphically, we mark the upper boundaries of the class intervals on the horizontal axis (X-axis) and their corresponding cumulative frequencies on the vartical axis (Y-axis), choosing a convenient scale. Now plot the points corresponding to the ordered pairs given by (upper boundary, corresponding cumulative frequency), i.e., (10, 5), (20, 8), (30,, 12), (40, 15), (50, 18), (60, 22), (70, 29), (80, 38), (90, 45), (100, 53) on a graph paper and join them by a free hand smooth curve. The curve we get is called a cumulative frequency curve, or an ogive (of the less than type).



The term 'ogive is pronounced as 'ojeev' and is derived from the word ogee. An ogee is a shape consisting of a concave arc flowing into a convex arc, so forming an S-shaped curve with vertical ends. In architecture, the ogee shape is one of the characteristics of the 14th and 15th century Gothic styles.

Again we consider the cumulative frequency distribution and draw its ogive (of the more than type).

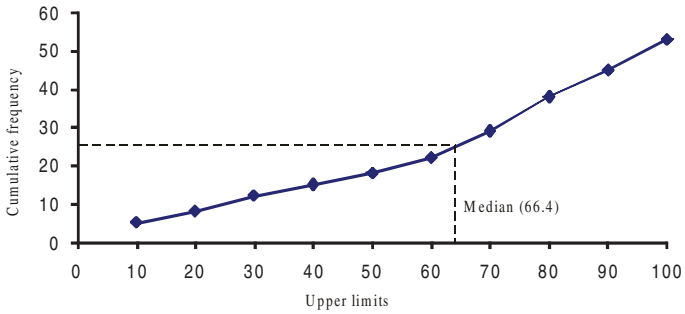


Recall that, here 0, 10, 20, ..., 90 are the lower boundaries of the respective class intervals 0-10, 10-20, ..., 90-100. To represent 'the more than type' graphically, we plot the lower boundaries on the X-axis and the corresponding cumulative frequencies on the Y-axis. Then we plot the points (lower boundaries, corresponding cumulative frequency), i.e., (0, 53), (10, 48), (20, (45), (30, 41), (40, 38), (50, 35), (60, 31), (70, 24), (80, 15), (90, 8), on a graph paper, and join them by a free hand smooth curve. The curve we get is a cumulative frequency curve, or an ogive (of the more than type).

14.6.1 Obtaining Median from give curve:

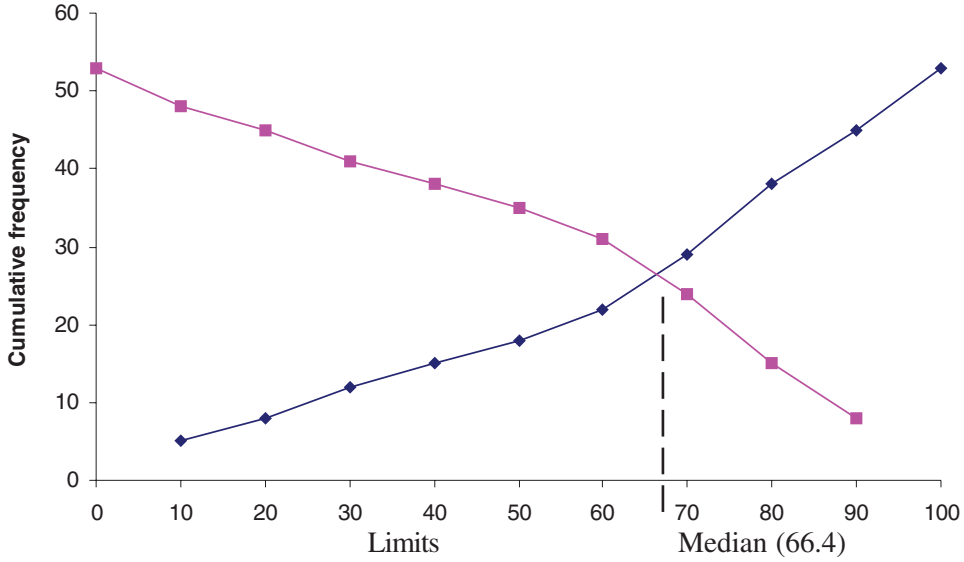
Is it possible to obtain the median from these two cumulative frequency curves . Let us see.

One obvious way is to locate on $\frac{n}{2} = \frac{53}{2} = 26.5$ on the y-axis. From this point, draw a line parallel to the x-axis cutting the curve at a point. From this point, draw a perpendicular to the x-axis. Foot of this perpendicular determines the median of the data.



Another way of obtaining the median :

Draw both ogives (i.e., of the less than type and of the more than type) on the same axis. The two ogives will intersect each other at a point. From this point, if we draw a perpendicular on the x-axis, the point at which it cuts the x-axis gives us the median.

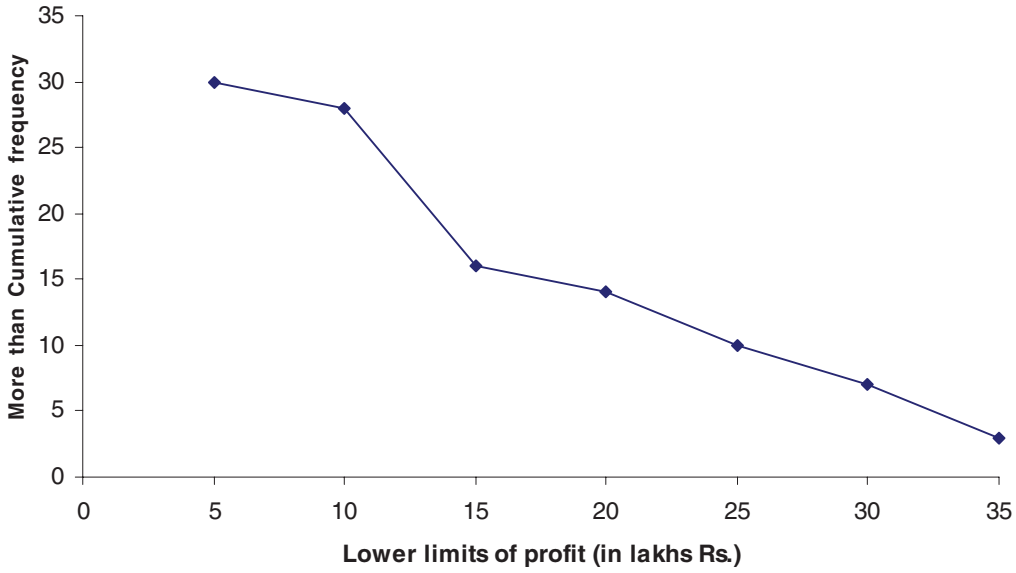


Example-9. The annual profits earned by 30 shops in a locality give rise to the following distribution :

Profit (in lakhs)	Number of shops (frequency)
More than or equal to 5	30
More than or equal to 10	28
More than or equal to 15	16
More than or equal to 20	14
More than or equal to 25	10
More than or equal to 30	7
More than or equal to 35	3

Draw both ogives for the data above. Hence obtain the median profit.

Solution : We first draw the coordinate axes, with lower limits of the profit along the horizontal axis, and the cumulative frequency along the vertical axes. Then, we plot the points (5, 30), (10, 28), (15, 16), (20, 14), (25, 10), (30, 7) and (35, 3). We join these points with a smooth curve to get the more than ovive, as shown in the figure below-

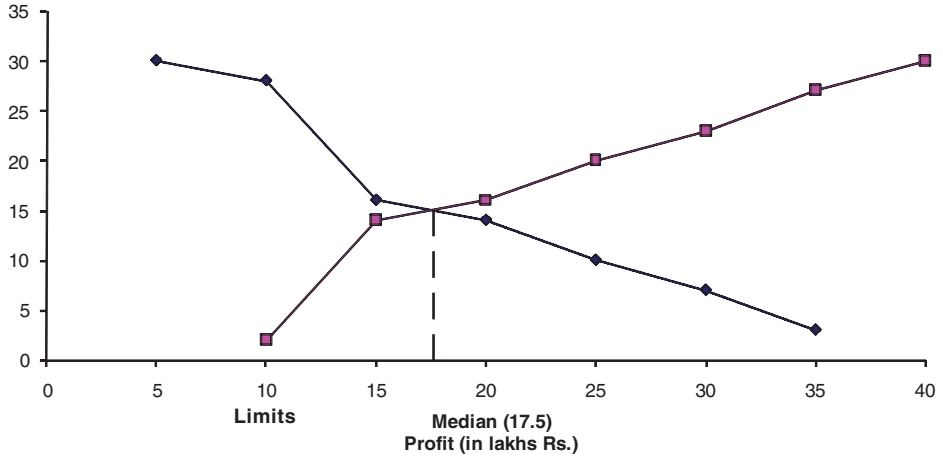


Now, let us obtain the classes, their frequencies and the cumulative frequency from the table above.

Classes	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Number of shops	2	12	2	4	3	4	3
Cumulative frequency	2	14	16	20	23	27	30

Using these values, we plot the points (10, 2), (15, 14), (20, 16), (25, 20), (30, 23), (35, 27), (40, 30) on the same axes as in last figure to get the less than ogive, as shown in figure below.

The abcissa of their point of intersection is nearly 17.5, which is the median. This can also be verified by using the formula. Hence, the median profit (in lakhs) is ₹ 17.5.



EXERCISE - 14.4

1. The following distribution gives the daily income of 50 workers of a factory.

Daily income (in Rupees)	250-300	300-350	350-400	400-450	450-500
Number of workers	12	14	8	6	10

Convert the distribution above to a less than type cumulative frequency distribution, and draw its ogive.

2. During the medical check-up of 35 students of a class, their weights were recorded as follows :

Weight (in kg)	Number of students
Less than 38	0
Less than 40	3
Less than 42	5
Less than 44	9
Less than 46	14
Less than 48	28
Less than 50	32
Less than 52	35

Draw a less than type ogive for the given data. Hence obtain the median weight from the graph and verify the result by using the formula.

3. The following table gives production yield per hectare of wheat of 100 farms of a village.

Production yield (Qui/Hec)	50-55	55-60	60-65	65-70	70-75	75-80
Number of farmers	2	8	12	24	38	16

Change the distribution to a more than type distribution, and draw its ogive.

What We Have Discussed

In this chapter, you have studied the following points :

1. The mean for grouped is calculated by :

   (i) The direct method : $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

   (ii) The assumed mean method : $\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$

   (iii) The step deviation method : $\bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \times h$

2. The mode for grouped data can be found by using the formula :

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

   where, symbols have their usual meaning.

3. The median for grouped data is formed by using the formula :

   $\text{Median} = l + \left(\frac{\frac{n}{2} - cf}{f}\right) \times h$ Where symbols have their usual meanings.

4. In order to find median, class intervals should be continuous.

5. Representing a cumulative frequency distribution graphically as a cumulative frequency curve, or an ogive of the less than type and of the more than type.

6. While drowing ogives boundaries are taken on X-main and cumulative frequencies are taken on Y-axis.

7. Scale on both the axes mya not be equal.

8. The median of grouped data can be obtained graphically as the x-coordinate of the point of intersection of the two ogives for this data.