5. Statistics

- **Construction of grouped frequency distribution table:**

  There are two ways to group the data to make frequency distribution table. These are as follows:

  Inclusive method (Discontinuous form):

  The classes can be defined in inclusive method as 1 - 10, 11 - 20, 21 - 30 and 31 - 40. Here, both limits are inclusive in each class.

  Exclusive method (Continuous form):

  In exclusive method, we take the class intervals as 0 – 10, 10 – 20, 20 – 30. The observations which are more than 0 but less than 10 will come under the group 0 – 10; the numbers which are more than 10 but less than 20 will come under the group 10 – 20 and so on. Here, the common observation will belong to the higher class, i.e. 10 will be included in the class interval 10 – 20 and similarly we follow this for the other observations also.

  For example, the ages of some residents of a particular locality are given as follows:

  7, 28, 30, 32, 18, 19, 37, 36, 14, 27, 12, 8, 17, 24, 22, 2, 21, 5, 21, 36, 38, 25, 10, 25, 9.

  Frequency distribution table can be drawn as follows:

.

  Inclusive method:

Class intervals	Tally marks	Frequency
1 – 10	𝍸 |	6
11 – 20	𝍸	5
21 – 30	𝍸 ||||	9
31 – 40	𝍸	5

  Exclusive method:

Class intervals	Tally marks	Frequency
1 – 10	𝍸	5
10 – 20	𝍸 |	6
20 – 30	𝍸 |||	8
30 – 40	𝍸 |	6

- **Few points to be remembered while choosing class intervals:**
  1. Classes should not be overlapping and all values or observations should be covered in these classes.
  2. The class size for all classes should be equal.
  3. The number of class intervals is normally between 5 and 10.

4. Class marks and class limits should be taken as integers or simple fractions.

1. The **central tendency** is the tendency of the numbers of a group to crowd around a certain number.

2. The three measures of central tendency are arithmetic mean, median and mode.

(i) The average number represents **Mean**.

(ii) The middle number represents **Median**.

(iii) The most often occurred number represents **Mode**.

3. Requirements of an **ideal measure of central tendency** are:

(i) It should be understood easily.

(ii) It should be calculated easily.

(iii) It should be based on all the observations of a data.

(iv) It should have one and only one interpretation i.e., it should be defined by a specific mathematical formula.

(v) It should be capable of further mathematical treatment.

(vi) It should be least affected by the extreme (end) observations in the data.

(vii) It should be determined using the graph of the data.

- **Mean of grouped data using direct method**

Mean $\overline{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$, where fi is the frequency corresponding to the class mark xi.

Example:

Consider the following distribution of marks scored by the students of a class in a unit test.

Marks scored	10 – 20	20 – 30	30 – 40	40 – 50
Number of students	4	7	15	14

Find the mean marks obtained by the students

Solution:

Class interval	**Frequency (f_i)**	**Class mark(x_i)**	$f_i x_i$
10 – 20	4	15	60
20 – 30	7	25	175
30 – 40	15	35	525
40 – 50	14	45	630
Total	$\Sigma f_i = 40$		$\Sigma f_i x_i = 1390$

Mean $= \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{1390}{40} = 34.75$

Thus, the mean of the marks obtained by the students is 34.75.

- **Assumed-mean method**

$\bar{x} = a + \bar{d} = a + \frac{\sum f_i d_i}{\sum f_i}$, where '$a$' is the assumed mean, $d_i = x_i - a$, and f_i is the frequency corresponding to the class mark x_i

Example:
The table below shows the attendance of students for 30 working days in a particular school.

Attendance	300 – 320	320 – 340	340 – 360	360 – 380	380 – 400
Number of days	8	6	7	6	3

Find the average attendance in this school.
Solution:

$$\textbf{Class marks} = \frac{\textbf{Upper limit} + \textbf{Lower limit}}{2}$$

$$\therefore x_1 = \frac{300+320}{2} = 310$$

$$x_2 = \frac{320+340}{2} = 330$$

$$x_3 = \frac{340+360}{2} = 350$$

$$x_4 = \frac{360+380}{2} = 370$$

$$x_5 = \frac{380+400}{2} = 390$$

Let the assumed mean 'a' be 350.

Class interval	**Number of days (f_i)**	**Class mark(x_i)**	$d_i = x_i - a$	$f_i d_i$
300 – 320	8	310	–40	–320
320 – 340	6	330	–20	–120
340 – 360	7	350 = a	0	0
360 – 380	6	370	+20	+120
380 – 400	3	390	+40	+120
Total	$\sum f_i = 30$			$\sum f_i d_i = -200$

$$\therefore \bar{x} = a + \frac{\sum f_i d_i}{\sum f_i} = 350 + \frac{(-200)}{30} = 350 - 6.67 = 343.33 \approx 343$$

Thus, the required average attendance in the school is 343 students per day.

- **Step-deviation method**

$$\bar{x} = a + h\bar{u} = a + h\left(\frac{\sum f_i u_i}{\sum f_i}\right), \text{where } u_i = \frac{x_i - a}{h},\ f_i$$

is the frequency corresponding to the class mark x_i , a is the assumed mean and h is the class size

Example: Find the mean of the following data.

Class interval	Frequency
600 – 800	4
800 – 1000	2
1000 – 1200	3
1200 – 1400	8
1400 – 1600	3

Solution:

Class size (h) = 200

Class interval	Frequency (f_i)	Class mark(x_i)	$d_i = x_i - a$	$u_i = \frac{x_i - a}{h}$	$f_i u_i$
600 – 800	4	700	–400	–2	–8
800 – 1000	2	900	–200	–1	–2
1000 – 1200	3	1100 = a	0	0	0
1200 – 1400	8	1300	200	1	8
1400 – 1600	3	1500	400	2	6
Total	20				4

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

$$= 1100 + 200 \times \frac{4}{20}$$

$$= 1100 + 40$$

$$= 1140$$

Thus, the required mean is 1140.

1. The assumed-mean method and the step-deviation method are simplified forms of the direct method
2. The mean obtained by all the three methods is the same.
3. Step-deviation method is convenient to apply if all d_i's have a common factor.

Note: If the class sizes are unequal, and x_i are numerically large, then the step-deviation method is still applicable by taking h to be suitable divisor of all the d_i's.

- **Median of grouped data**

  Median of a grouped data is given by:

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

  Median

where l = Lower limit of median class
n = Number of observations
cf = Cumulative frequency of the class preceding the median class
f = Frequency of the median class
h = Class size (assuming class size to be equal)

Example: Find the median of the following distribution.

Class interval	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100	100 – 120
Frequency	7	8	6	8	6	5

Solution: The cumulative frequency for the given data can be written as:

Class interval	**Frequency**	**Cumulative frequency**
0 – 20	7	7
20 – 40	8	7 + 8= 15
40 – 60	6	15 + 6 = 21
60 – 80	8	21 + 8 = 29
80 – 100	6	29 + 6 = 35
100 – 120	5	35 + 5 = 40

Here, $n = 40$

$$\therefore \frac{n}{2} = \frac{40}{2} = 20$$

lies in the class 40 – 60

Median class is 40 – 60

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

$$l = 40, cf = 15, f = 6, h = 20$$

$$\therefore \text{Median} = 40 + \left(\frac{20 - 15}{6} \right) \times 20$$

$$= 40 + \frac{5}{6} \times 20$$

$$= 40 + 16.66 (\text{approx})$$

$$= 55.66 (\text{approx})$$

1. Merits of mean:

(i) It is very simple to understand.

(ii) It is very easy to calculate.

(iii) It is based on all the observations of a series.

(iv) It can be determined for almost every kind of data.

(v) It is capable of further mathematical treatment.

(vi) Mean has one and only one interpretation as it is defined by a specific mathematical formula.

(vii) It is least affected by fluctuation of data as it depends on all observations involved in the data.

2. Demerits of mean:

(i) It cannot be determined by just inspecting the data.

(ii) It cannot be determined from the graph of the data.

(iii) It cannot be computed for frequency distributions having open end classes.

(iv) It is too much affected by the extreme (end) observations in the data.

(v) It is not appropriate in case of highly asymmetric data.

(vi) It does not give accurate measure of central tendency for ratios and percentages.

(vii) It cannot be determined if any item or value is missing from data.

1. Merits of median:

(i) It is very simple to understand.

(ii) It can be easily calculated.

(iii) It can be determined by just observation in some cases.

(iv) It can be determined from graph of the data.

(v) It is not affected by change in extreme values.

2. Demerits of median:

(i) Median of a data is incapable of further algebraic or mathematical treatment.

(ii) Median is an approximate value, not a precise value of a series as when a series has even number of terms, the median is located somewhere between the two middle values.

(iii) If the data contains large number of observations, the process of arranging the observations in ascending or descending order for calculating median becomes tiresome.

(iv) Median cannot represent all the items of a data set.

(v) Median is very much affected by fluctuations in the data.

- **MODE**
    - **Mode of ungrouped data**

The mode or modal value of a distribution is the observation for which the frequency is the maximum.

- **Mode of grouped data**

Mode of a grouped data is given by:

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

where, l = Lower limit of the modal class
h = Size of the class interval (assuming all class sizes to be equal)
f_1 = Frequency of the modal class
f_0 = Frequency of the class preceding the modal class
f_2= Frequency of the class succeeding the modal class

Example: Find the mode of the following distribution.

Class interval	Frequency
0 – 5	4
5 – 10	9
10 – 15	7
15 – 20	10
20 – 25	5
25 – 30	6

Solution: The maximum class frequency is 10.
Modal class is 15 – 20
$l = 15, h = 5$
$f_1 = 10, f_0 = 7, f_2 = 5$

$$\therefore \text{Mode} = 15 + \left(\frac{10-7}{2\times10-7-5} \right) \times 5$$

$$= 15 + \frac{15}{8}$$

$$= 15 + 1.875 = 16.875$$

1. Merits of mode:

(i) It is very simple to understand.

(ii) It can be easily calculated.

(iii) It is not affected by the extreme values.

(iv) When the data is ungrouped, it can be obtained just by inspecting the data.

(v) It can also be determined by drawing graphs.

(vi) It can be computed for frequency distributions having open end classes.

2. Demerits of mode:

(i) It is independent from other observations.

(ii) It cannot be used for further mathematical treatments.

(iii) It is not unique.

(iv) It may or may not exist.

(v) It cannot be determined if the modal class lie at the end of the distribution in case of grouped frequency distribution.

- **Empirical relationship between the three measures of central tendency**

  3 Median = Mode + 2 Mean

- **Pie chart**

  A pie chart or a circle graph shows the relationship between a whole and its parts.

- **Construction of pie charts**

  Example:

  Construct a pie chart for the following data which gives the brands of laptop preferred by the people of a locality.

  Brand A : 100

  Brand B : 120

  Brand C : 180

  Solution:

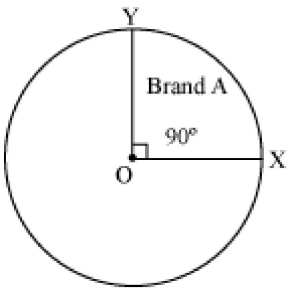
  The total number of people is 100 + 180 + 120 = 400.

  We can form the following table to find the central angle of each sector:

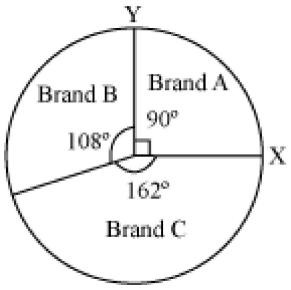
Brand of laptop	Number of people	Fraction	Central angle
A	100	$\frac{100}{400} = \frac{1}{4}$	$\frac{1}{4} \times 360° = 90°$
B	180	$\frac{120}{400} = \frac{3}{10}$	$\frac{3}{10} \times 360° = 108°$
C	120	$\frac{180}{400} = \frac{9}{20}$	$\frac{9}{20} \times 360° = 162°$

  Steps of construction:

  - Draw a circle with any convenient radius. Let O be the centre of the circle and OX be its radius.
  - Draw the angle of the sector for brand A, which is 90°. Using protractor, draw $\angle XOY = 90°$.



- Now, draw the angle of the sectors for brands B and C.

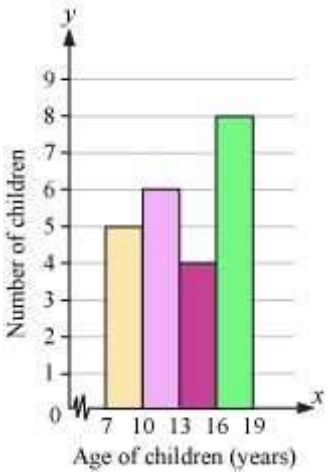


- **Histogram**

  A histogram is a bar graph that is used to represent grouped data. In a histogram, the class intervals are represented on the horizontal axis and the heights of the bars represent frequency. Also, there is no gap between the bars in a histogram.

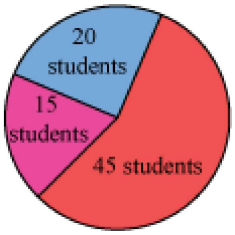
Class interval (Age of children)	**Tally mark**	**Frequency (Number of children)**
7 – 10	𝍸	5
10 – 13	𝍸 |	6
13 – 16	||||	4
16 – 19	𝍸 |||	8

The above frequency distribution table can be displayed in a histogram as follows:



In a histogram, a broken line can be used along the horizontal axis to indicate that the numbers between 0 to7 are not included.

- For example: Consider the given pie chart which shows the favourite colours of the class-VIII students of a school.



In this pie chart, the portion of the sector for the colour red is given by,

$$\frac{\text{Number of students whose favourite colours is red}}{\text{Total number of students}}$$

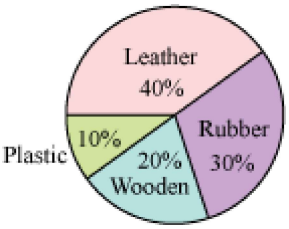
$$= \frac{45}{80}$$

$$= \frac{9}{16}$$

Therefore, the sector representing red colour is $\left(\frac{9}{16}\right)^{th}$ part of the circle.

- **Interpretation of a pie chart**

  The given pie chart shows the footwears preferred by the people of a locality.



From the above pie chart, we can infer that most people of the locality prefer wearing leather footwears. Also, we can infer that the least number of people prefer wearing plastic footwears.

Now, suppose the total number of people in the locality is 1000. Then, we can say that the number of people who prefer wearing rubber footwears is 30100×1000 = 300

1. Uses of histogram:

(i) Histograms are used to find mode of the data.

(ii) They are helpful in the construction of frequency curve and frequency polygon.

(iii) They are useful to figure out the location of average.

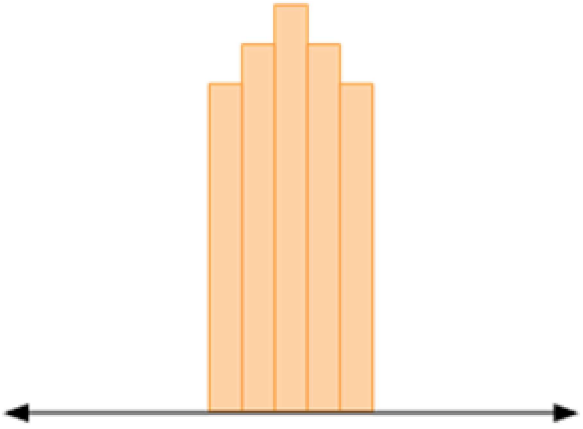
(iv) They give information about the spread of the data.

(v) These are helpful to get the idea about the symmetry of frequency distribution.

(vi) In case of mixture of two data sets, histogram is bimodal in nature.

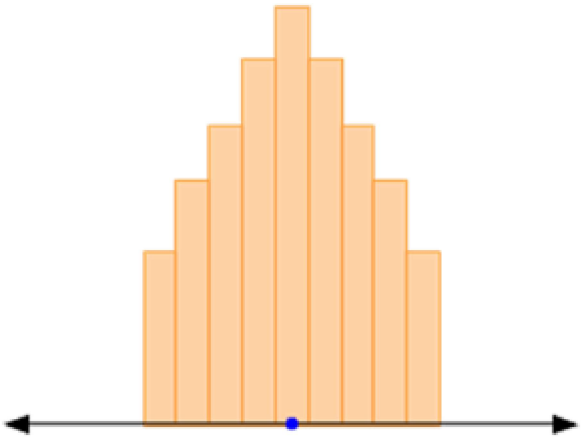
2. Histograms of different types of data:

(i) Narrow spreaded distribution:

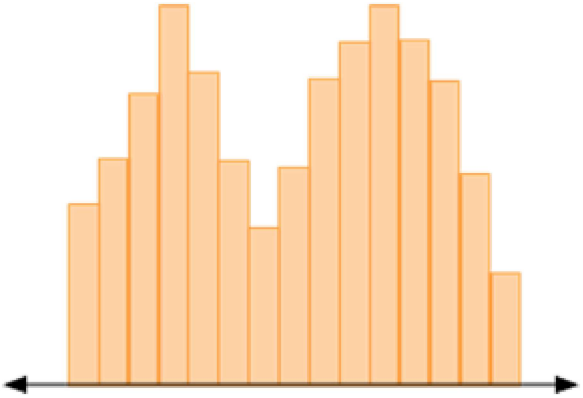


(ii) Widely spreaded distribution:

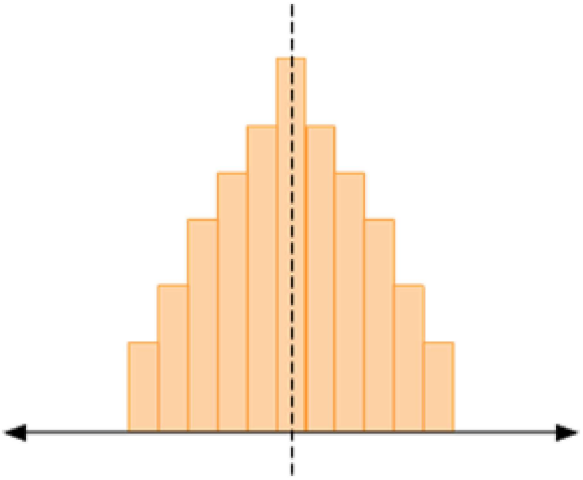
(iii) Average distribution:



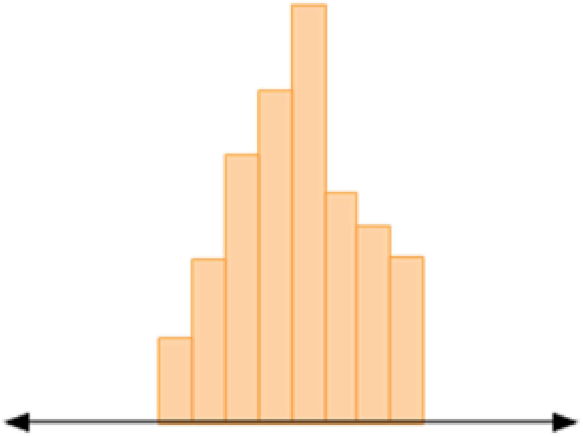
(iv) Bimodal distribution:



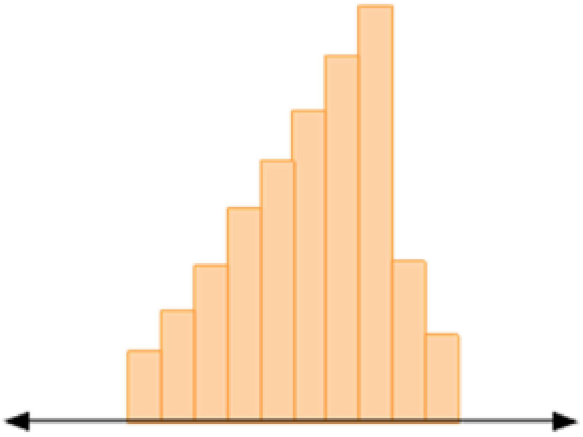
(v) Perfectly symmetric distribution:



(vi) Asymmetric distribution:



(vii) Highly asymmetric distribution:



- The observation with maximum frequency is called **mode**. When data is grouped into classes, mode can be obtained by drawing histogram.

  Example:

  The following table shows the class intervals and the frequency corresponding to them.

Class Interval	40 − 69	70 − 99	100 − 129	130 − 159	160 − 189
Frequency	12	15	24	16	21

Find the mode of the given data geometrically.

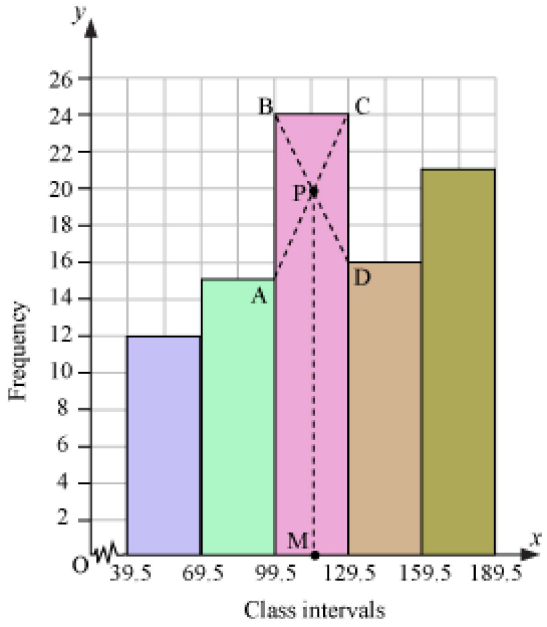
Solution:

The given frequency distribution is discontinuous. To convert it into continuous distribution, we have to subtract $\frac{1}{2} = 0.5$ from the lower limits and add 0.5 to higher limits of each class interval.

Now, the continuous frequency distribution table of the given data is as follows:

Class Interval	Frequency
39.5 − 69.5	12
69.5 − 99.5	15
99.5 − 129.5	24
129.5 − 159.5	16
159.5 − 189.5	21

To find the mode of the above data geometrically, first of all we have to draw its histogram by choosing 1 cm along *x*-axis = 30 (class-intervals) and 1 cm = 2 (frequencies). In the highest rectangle (class interval 99.5 − 129.5), we will draw two straight lines AC and BD from corners of the rectangles on either side of the highest rectangle to the opposite corners of the highest rectangle. Let P be the intersection of the lines AC and BD. Now, we will draw a vertical line through the point P that cuts the *x*-axis at M.



The point M represents the value 115 (approximately) on *x*-axis. Therefore, the mode of the given data is 115 (approximately).

- **Construction of frequency polygons**

A frequency polygon is a continuous curve obtained by plotting and joining the ordered pairs of class marks and their corresponding frequencies.
There are two ways to construct a frequency polygon.

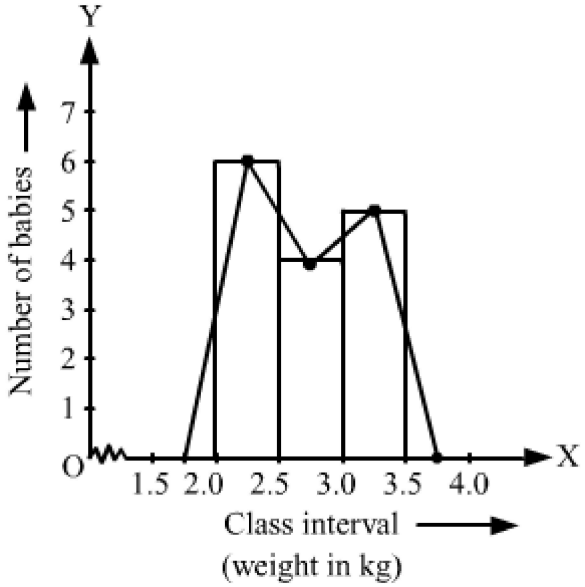
- The frequency polygon for a grouped data is drawn by first drawing its histogram and then by joining the mid-points of the top of bars and the mid-points of the classes preceding and succeeding the lowest and highest class respectively.
- One other way of drawing a frequency polygon is by plotting and joining the ordered pairs (of class marks and their corresponding frequencies) with the mid-points of the classes preceding and succeeding lowest and highest class respectively.

Example:
Here are the weights (in kg) of the babies born in a hospital during a particular week.
2.3, 2.0, 2.5, 2.7, 3.0, 3.2, 3.1, 2.2, 3.0, 2.5, 2.4, 3.0, 2.3, 2.4, 2.8
Draw a histogram for the data and then draw a frequency polygon using it.

Solution:
The frequency distribution table of the given data is as follows:

Class interval	Frequency
2.0–2.5	6
2.5–3.0	4
3.0–3.5	5

The histogram and frequency polygon for the given data can be drawn as:



- **Graphical representation of cumulative frequency distribution Ogive**
  - **OGIVE (of the less- than type)**

Example 1: Draw ogive of the less-than type for the given distribution.

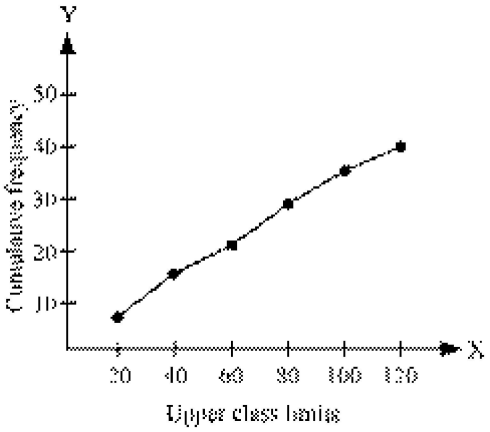
Class interval	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100	100 – 120

Frequency	7	8	6	8	6	5

Solution: The cumulative frequency distribution for the given data can be found as:

Class interval	Upper class limit	Frequency	Cumulative frequency
0 – 20	20	7	7
20 – 40	40	8	15
40 – 60	60	6	21
60 – 80	80	8	29
80 – 100	100	6	35
100 – 120	120	5	40

By taking the horizontal axis as the upper class limit and the vertical axis as the corresponding cumulative frequency, we can plot the cumulative frequency for each upper class limit.
Then, the required ogive (of the less-than type) is obtained as:



- **OGIVE (of the more-than type)**

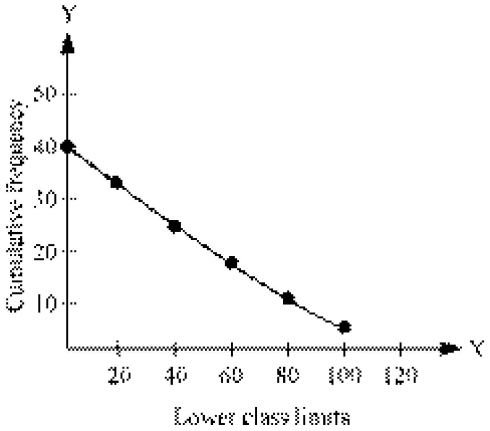
Example 2: Draw ogive of the more-than type for the following distribution.

Class interval	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100	100 – 120
Frequency	7	8	6	8	6	5

Solution: The cumulative frequency for the given data can be found as:

Class interval	Lower class limit	Frequency	Cumulative frequency
0 – 20	0	7	40
20 – 40	20	8	33
40 – 60	40	6	25
60 – 80	60	8	19
80 – 100	80	6	11
100 – 120	100	5	5

By taking the horizontal axis as the lower class limit and the vertical axis as the corresponding cumulative frequency, we can plot the cumulative frequency for each lower class limit.
Then, the required ogive (of the more-than type) is obtained as:



Note:

The *x*-coordinate of the point of intersection of the "more-than ogive" and "less-than ogive" of a given grouped data gives its median.