
UNIT 2 DESCRIBING DATA SETS AND MEASURES OF CENTRAL TENDENCY

Structure

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Frequency Distribution
- 2.3 Grouping Data in Class Intervals
- 2.4 Graphical Representation of Data
 - 2.4.1 Graphs of Frequency Distributions
 - Histogram
 - Frequency Polygon
 - Frequency Curve
 - Cumulative Frequency Curves - Ogive
 - 2.4.2 Stem and Leaf Plots
- 2.5 Measures of Central Tendency
 - 2.5.1 Mean
 - 2.5.2 Median
 - 2.5.3 Mode
 - 2.5.4 Relationship between Mean, Median and Mode
 - 2.5.5 Geometric and Harmonic Mean
- 2.6 Partition Values
 - 2.6.1 Quartiles
 - 2.6.2 Deciles
 - 2.6.3 Percentiles
 - 2.6.4 Boxplots
- 2.7 Summary
- 2.8 Solutions /Answers
- 2.9 References

2.0 INTRODUCTION

In Unit 1 of this block, we have defined the concept of statistics, its uses, types of data, data creation, etc. In this unit, we present how you can tabulate and represent data graphically. In this unit, we are going to discuss about the basic analysis of data. We first define the frequency distribution, which aims at classifying measurable data. This data can then be represented using graphs that communicate the characteristics of the data more effectively. Further, for the purpose of analysis, a very important and powerful tool is a single average value that represents the entire mass of data.

The term average in Statistics refers to a one figure summary of a distribution. It gives a value around which the distribution is concentrated. For this reason, the average is also called the measure of central tendency. The calculation of the average condenses a distribution into a single value that is supposed to represent the distribution. This helps both individual assessments of a distribution as well as in comparison with another distribution.

This unit comprises of frequency distribution and its graphical representations. It various measures of central tendency and concepts and methods of calculating the partition values.

2.1 OBJECTIVES

After studying this unit, you would be able to

- define different types of frequency distribution;
 - make graphs like histogram, ogive, stem and leaf plot;
 - explain the measure of central tendency;
 - calculate the different types of measures of central tendency;
 - describe the methods of calculation of partition values;
 - make a boxplot using the partition values.
-

2.2 FREQUENCY DISTRIBUTION

A frequency distribution refers to the data which are classified on the basis of some variables that can be measured such as wages, age of children, etc. A variable refers to the characteristic that varies in magnitude in a frequency distribution. It may be either discrete or continuous. A discrete variable is one that generally takes integer values. For example, the number of students, the number of books, etc. A continuous variable can take integer or fractional values within the range of possibilities, such as the height or weight of individuals. Generally speaking, continuous data are obtained through measurements while discrete data are derived by counting. A series described by a continuous variable is called continuous series. Similarly, series represented by a discrete variable is called discrete series.

According to the nature of the variable, the frequency distribution may be of two types, i.e. discrete frequency distribution and continuous frequency distribution.

Discrete Frequency Distribution: A frequency distribution in which the information is distributed in different classes on the basis of a discrete variable is known as a discrete frequency distribution.

Continuous Frequency Distribution: A distribution in which the information is distributed in different classes on the basis of a continuous variable is known as continuous frequency distribution. There may be some variables which have integer values as well as fractional values. Frequency distribution of such variables is called continuous frequency distribution.

Let us discuss the Relative and Cumulative frequency distributions which are of the similar importance as analysis point of view of data is considered.

Relative Frequency Distribution

A relative frequency corresponding to a class is the ratio of the frequency of that class to the total frequency. The corresponding frequency distribution is called relative frequency distribution. If we multiply each relative frequency by 100, we get the percentage frequency corresponding to that class and the corresponding frequency distribution is called “Percentage frequency distribution”. Let us take an example in which both relative and percentage frequency distributions are prepared.

Example 1: A frequency distribution of marks of 50 students in a subject is as given below:

Class (Marks): 0-10 10-20 20-30 30-40 40-50

Frequency: 6 10 14 18 2

Prepare relative and percentage frequency distributions.

Solution: The relative and percentage frequency distributions can be formed as given in the following table:

Class (Marks) X	Frequency (f)	Relative frequency (f/N)	Percentage Frequency (f/N) × 100
0-10	6	$6/50 = 0.12$	$0.12 \times 100 = 12\%$
10-20	10	$10/50 = 0.20$	$0.20 \times 100 = 20\%$
20-30	14	$14/50 = 0.28$	$0.28 \times 100 = 28\%$
30-40	18	$18/50 = 0.36$	$0.36 \times 100 = 36\%$
40-50	2	$2/50 = 0.04$	$0.04 \times 100 = 4\%$
Total	$\sum f = N = 50$	1.00	100

Cumulative Frequency Distribution

The cumulative frequency of a class is the total of all the frequencies up to and including that class. A cumulative frequency distribution is a frequency distribution which shows the observations 'less than' or 'more than' a specific value of the variable.

The number of observations less than the upper class limit of a given class is called the less than cumulative frequency and the corresponding cumulative frequency distribution is called less than cumulative frequency distribution. Similarly, the number of observations corresponding to the value of more than the lower class limit of a given class is called more than cumulative frequency and the corresponding cumulative frequency distribution is called 'more than' cumulative frequency distribution. Following is an example, wherein 'less than' and 'more than' cumulative frequency distributions have been obtained.

Example 2: For the following frequency distribution of marks of 50 students in a subject, form both types of cumulative frequency distributions.

Class (Marks)	0-10	10-20	20-30	30-40	40-50
No. of Students	7	11	15	12	5

Solution: Cumulative frequency distributions are formed as given in the following table:

Given Frequency Distribution		Less Than Cumulative Frequency Distribution		More Than Cumulative Frequency Distribution	
Classes	No. of Students	Marks Less than	No. of students	Marks More than	No of students
0-10	07	10	07	0	50
10-20	11	20	18	10	43
20-30	15	30	33	20	32
30-40	12	40	45	30	17
40-50	05	50	50	40	05
Total	50				

After discussing the frequency distributions, we now discuss how the concept of frequency distribution can be used to classify the data according to the class intervals in the next section.

2.3 GROUPING DATA IN CLASS INTERVALS

To make data understandable, data are divided into number of homogeneous groups or subgroups. In classification, according to class intervals, the observations are arranged systematically into a number of groups called classes. Such classification is most popular in practice. But before this discussion we have to define some terms which will be used in the above classification.

- (i) **Class Limits:** The class limits are the lowest and the highest values of a class. For example, let us take the class 10-20. The lowest value of this class is 10 and the highest 20. The two boundaries of a class are known as the lower limit and upper limit of the class.
- (ii) **Class Intervals:** The class interval of a class is the difference between the upper class limit and the lower class limit. For example, in the class 10-20 the class interval is 10 (i.e. 20 minus 10). This is valid in the case of exclusive method discussed in this section later. If the inclusive frequency distribution (discussed later in this section) is given then first it is converted to exclusive form and then class interval is calculated. The size of the class interval is determined by number of classes and the total range of data.
- (iii) **Range of Data:** The range of data may be defined as the difference between the lower class limit of the first class interval and the upper class limit of the last class interval.
- (iv) **Class Frequency:** The number of observations corresponding to the particular class is known as the frequency of that class or the class frequency. If we add together the frequencies of all individual classes, we obtain the total frequency.
- (v) **Class Mid Value:** It is the value lying halfway between the lower and upper class limits of class-interval, mid-point or mid value of a class is defined as follows:

$$\text{Mid Value of a Class} = \frac{\text{Upper class limit} + \text{Lower class limit}}{2}$$

For the purpose of further calculations in statistical analysis, mid value of each class is taken to represent that class.

Now we are in position to discuss the two methods of classification according to class intervals, namely “**Exclusive Method**” and “**Inclusive Method**”. Let us discuss these two methods one by one:

Exclusive Method

Under this method, a class interval is such that each upper class limit is excluded from the class interval. Here, in this method, class intervals are so fixed that the upper limit of one class is the lower limit of the next class. For example, in exclusive method a student who has secured 20 marks would be included in class 20-30, and not in 10–20. This method is widely followed in practice.

Example 3: 24 students appeared in an entrance test where all questions are objective type with 25% negative marking. The marks obtained out of 50 maximum marks are as follows:

17, 16, 7, 30, 21, 42, 44, 36, 22, 22, 25, 31, 31, 34, 30, 36, 35, 45, 25, 15, 20, 42, 40, 30

Prepare a frequency distribution by using exclusive method.

Solution: Frequency distribution of marks obtained by above 24 students is given below in Table 2.1 using the exclusive method.

Table 2.1: Frequency Distribution of 24 Students by Exclusive Method

Classes	Tally bar	No. of Students
0-10		1
10-20		3
20-30		6
30-40		9
40-50		5
Total		24

Inclusive Method

Under the inclusive method of classification both lower class limit as well as the upper limit of a class are included in that class itself. Following frequency distribution is formed using inclusive method for the data of Example 3 given above.

Table 2.2: Frequency Distribution of 24 Students by Inclusive Method

Class	Tally bar	No. of Students
0-9		1
10-19		3
20-29		6
30-39		9
40-49		5
Total		24

For converting data from *inclusive form* to *exclusive form*, first of all we find the half of the difference of lower limit of that class and upper limit of the preceding class. This value is then subtracted from lower limit of each class and added to the upper limit of each class. In the above example, this can be easily understood as $(10-9)/2 = 0.5$. So, the class intervals are as – 0.5- 9.5, 9.5-19.5, ..., 39.5-49.5. If all the observations of data are positive, then the lower limit of the first class can be taken as 0. Therefore, in this case the class intervals are as 0-9.5, 9.5-19.5, ..., 39.5-49.5.

Remark

- Lower limit of a class interval is always included in the class in both the method discussed above.
- In exclusive method upper limit of a class is not included in the class. That is why the name exclusive.

- (iii) In inclusive method upper limit of a class is also included in the class. That is why the name inclusive.

Principles of Classification

It is difficult to formulate any hard and fast rule for classifying the data. However, the following general considerations may be considered for ensuring meaningful classification of data:

- (1) The whole data should be preferably divided into number of classes between 5 and 15. To determine the approximate number of classes (K) the following formula is suggested by “Struges”:

$$K = 1 + 3.322 \text{ Log } N, \text{ where } K = \text{the approximate number of classes}$$

N = total number of observations

Log = the natural logarithm

However, the appropriate number of classes to be taken for a given data depends upon the personal judgment and other considerations such as range of data, total number of observations, etc.

- (2) One should avoid odd values of class intervals as far as possible, e.g. 3, 7, 11, 26, 39, etc. One should prefer 5 or 10 or multiple of 5 or 10 as class intervals such as 5, 10, 20, 25, 100, etc, because the human mind is accustomed more to think in terms of certain multiples of 5 or 10.
- (3) The lower class limit of the first class of a frequency distribution should either be zero or 5 or multiple of five. For example if the lowest value of the data is 26 and we have taken a class interval of 5, then the first class should be 25-30, instead of 26-31.
- (4) To maintain continuity and to get correct class interval, we should adopt exclusive method of classification. However, where ‘inclusive’ method has been adopted it is necessary to make an adjustment to determine the correct class interval and to maintain continuity.
- (5) The intervals of all the classes should be of the same size, because if the class intervals are not of the same width, it is difficult to make meaningful comparisons between classes. Sometimes the data may require the inclusion of so many class intervals that the frequency distribution will become large. Then the classification may be done as follows:

below 10; 10-20; 20-30; 30-40; above 40

These classes are called open ended classes and the distribution is known as open ended frequency distribution.

It may be noted that the frequency distributions, like other types of data presentation, are always constructed to serve some specific purpose. The technical requirements outlined above must be supplemented by sound subjective judgments if proper frequency distributions are to be formed.

Check Your Progress 1

- 1 The marks of 30 students in statistics are given below:

10, 12, 25, 32, 27, 32, 38, 43, 39, 55, 29, 38, 57, 08, 06, 13, 27, 25, 29, 53, 55, 45, 35, 48, 47, 59, 15, 19, 48, 55

Classify the above data by taking a suitable class interval.

.....

.....

- 2 Present the following data of the profits (in crores of Rs.) of the 60 companies in the years 2025-26:

41, 17, 83, 63, 55, 92, 60, 58, 70, 06, 67, 82, 33, 44, 57, 49, 34, 73, 54,
63, 36, 52, 32, 75, 60, 33, 09, 79, 28, 30, 42, 93, 43, 80, 03, 32, 57, 67,
84, 64, 63, 11, 35, 28, 10, 23, 08, 41, 60, 32, 72, 53, 92, 88, 62, 55, 60,
33, 40, 57

Classify data by inclusive method.

.....

.....

- 3 Use the data given in the question 2 to present the same using principle of adding and subtracting the correction factor.
-
-

2.4 GRAPHICAL REPRESENTATION OF DATA

A graphical presentation is a geometric image of a set of data. Graphical presentation is done for both frequency distributions and time series. One of the important features of graphs is that if a person once sees the graphs, the figure representing the graphs is kept in his/her brain for a long time. They also help us in studying cause and effect relationship between two variables. The graph of a frequency distribution presents the huge data in an interesting and effective manner and brings to light the salient features of the data at a glance. Let us see some advantages of graphical presentation.

Advantages of Graphical Presentation

The following are some advantages of the graphical presentation:

- It simplifies the complexity of data and makes it readily understandable.
- It attracts attention of people.
- It saves time and efforts to understand the facts.
- It makes comparison easy.
- A graph describes the relationship between two or more variables.

Now a days a large variety of graphs are in practical use. However, we shall discuss only some important graphs which are more popularly used in practice. The various types of graphs can be divided broadly under the following two heads:

- Graphs of Frequency Distributions
- Graphs for Time Series Data

In this Unit, we will focus on graphs for frequency distributions.

2.4.1 Graphs of Frequency Distributions

The graphical presentation of frequency distributions is drawn for discrete as well as continuous frequency distributions.

Let us first consider the frequency distribution of a discrete variable.

Frequency Bar Diagram: To represent a discrete frequency distribution graphically, we take two rectangular axes of coordinates, the horizontal axis for the variable and the vertical axis for the frequency. The different values of the variable are then located as points on the horizontal axis. At each of these points, a perpendicular bar is drawn to present the corresponding frequency.

Such a diagram is called a 'Frequency Bar Diagram'. For example, if we take the frequency distribution for the number of peas per pod for 198 pods as given in Table 2.3:

Table 2.3: A Sample Frequency Distribution

No of peas per pod	1	2	3	4	5	6	7
Frequency (number of pods)	14	23	66	40	26	18	11

Then, the frequency bar diagram is shown in Fig. 2.1:

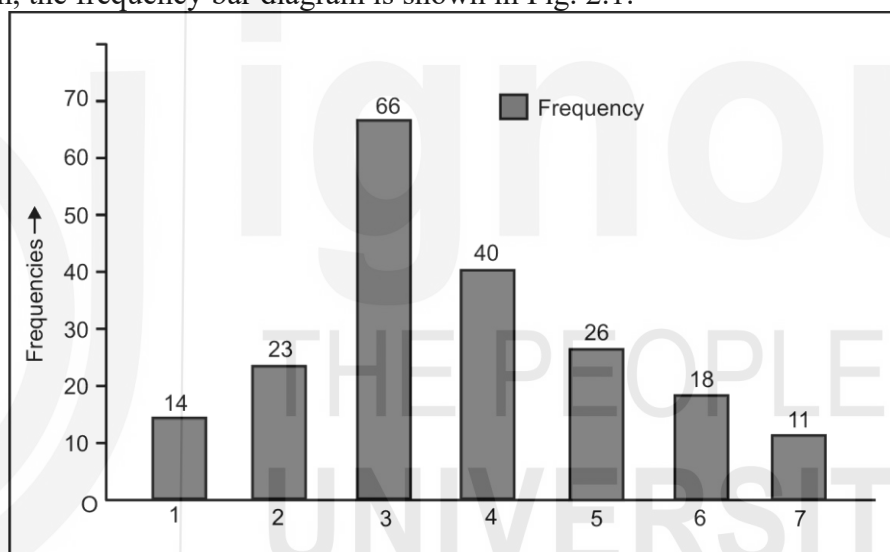


Fig. 2.1: Frequency Bar Diagram for the Frequency Distribution of Number of the Peas for 198 Pods.

Note 1: O represents origin and choice of scale used along horizontal and vertical axes depends upon given data.

Now, we take the case of frequency distribution of a continuous variable.

The following are the most commonly used graphs for continuous frequency distributions:

- (i) Histogram
 - (ii) Frequency Polygon
 - (iii) Frequency Curve
 - (iv) Cumulative Frequency Curve or Ogives
- Let us discuss these one by one:

(i) Histogram

In previous example, we have discussed how a graph is drawn for discrete frequency distribution.

For the continuous frequency distribution, a better way to represent the data graphically is to use a histogram. A histogram is drawn by constructing adjacent rectangles over the class intervals such that the length of the rectangles is proportional to the corresponding class frequencies.

Histogram is similar to a bar diagram which represents a frequency distribution with continuous classes. The width of all bars is equal to class interval. Each rectangle is joined with the other so as to give a continuous picture.

The class-boundaries are located on the horizontal axis. If the class-intervals are of equal size, the heights of the rectangles will be proportional to the class-frequencies themselves. If the class-intervals are not of equal size, the heights of the rectangles will be proportional to the ratios of the frequencies to the width of the corresponding classes. In other words, the frequencies of the class-intervals having the least width are written as they are and the frequencies of other class intervals are written as follows:

$$\frac{\text{Given frequency}}{\text{Width of its Class-interval}} \times (\text{The least width}) \quad \dots (1)$$

Let us draw a histogram to the following frequency distribution given below in the Table 2.4.

Table 2.4: A Sample Frequency Distribution for Class Intervals of Equal Width

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	2	3	13	18	9	7	6	2

Histogram for the above data is given below.

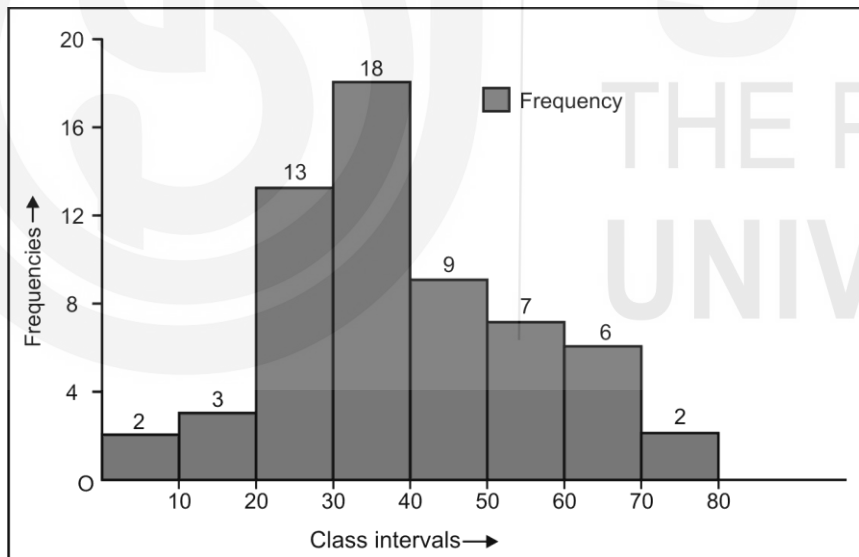


Fig. 2.2: Histogram for Frequency Distribution when Class-intervals are of Equal Width.

Now, let us consider the frequency distribution for unequal class intervals as given in Table 2.5

Table 2.5: A Frequency Distribution for Class Intervals of Unequal Width

Class	0-10	10-20	20-30	30-40	40-70	70-80	80-100
Frequency	20	32	8	2	60	35	10

As it is a case of unequal class intervals, so we have to adjust the frequencies of the classes 40-70 and 80-100 by the formula suggested in equation 1. These calculations are shown in Table 2.6 given below:

Table 2.6: Computation of Height of Rectangles in Histograms

Class Interval (CI)	Frequency	Width of (CI)	Heights of the rectangles
0-10	20	10	20
10-20	32	10	32
20-30	8	10	8
30-40	2	10	2
40-70	60	30	$(60/30) \times 10 = 20$
70-80	35	10	35
80-100	10	20	$(10/20) \times 10 = 5$

The histogram for this frequency distribution is shown in Fig. 2.3.

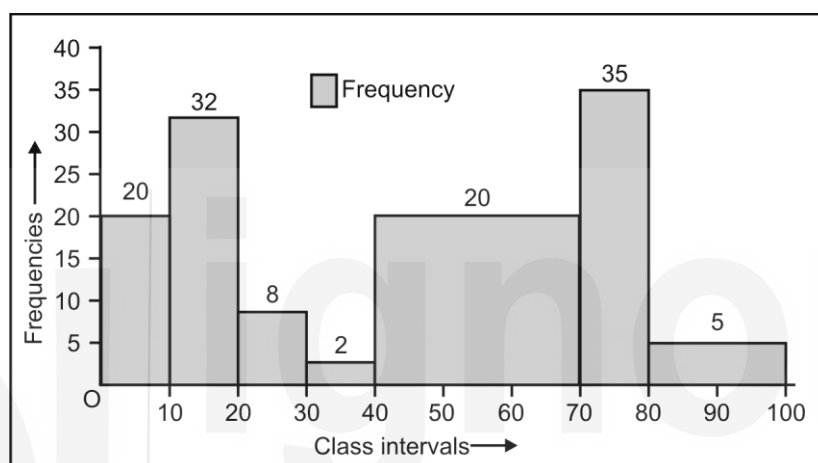


Fig. 2.3: Histogram for Frequency Distribution when Class Intervals are of Unequal Width

Note 2: Sometimes, a histogram is also used for the frequency distribution of a discrete variable. Each value of the discrete variable is regarded as the mid-point of an interval. But generally, its use is not recommended, because in discrete case each frequency actually corresponds to a single point and not to an interval.

(ii) Frequency Polygon

Another method of presenting a frequency distribution graphically other than histogram is to use a frequency polygon. In order to draw the graph of a frequency polygon, first of all the mid values of all the class intervals and the corresponding frequencies are plotted as points with the help of the rectangular coordinate axes. Secondly, we join these plotted points by line segments. The graph thus obtained is known as frequency polygon, but one important point to keep in mind is that whenever a frequency polygon is required, we take two imaginary class intervals each with frequency zero, one just before the first class interval and other just after the last class interval. Addition of these two class intervals facilitate the existence of the property that

$$\text{Area under the polygon} = \text{Area of the histogram}$$

For example, if we take the frequency distribution as given in Table 2.4, then we have to first plot the points (5, 2), (15, 3), ..., (75, 2) on graph paper along with the horizontal bars. Then we join the successive points (including the midpoints of two imaginary class intervals, each with zero frequency) by line segments to

get a frequency polygon. The frequency polygon for frequency distribution given in Table 2.4 is shown in Fig. 2.4.

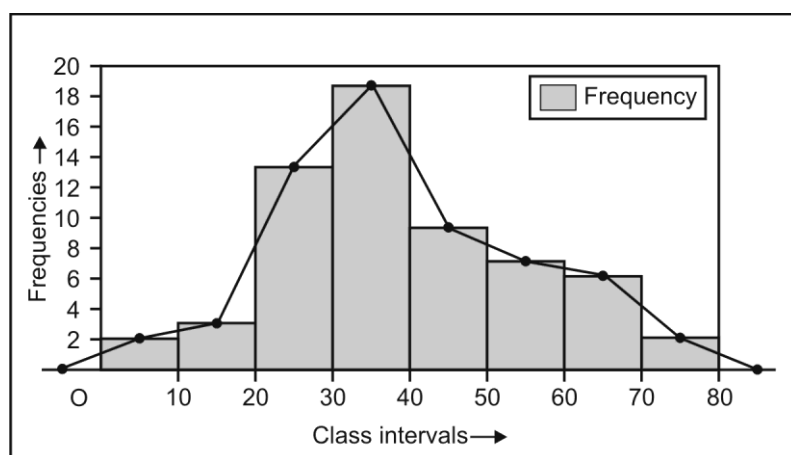


Fig. 2.4: Frequency Polygon for the Frequency Distribution given in Table 2.4.

Note 3: In some cases, the first class interval does not start from zero. In such situations, we mark a kink on the horizontal axis, which will indicate the continuity of the scale starting from zero. Let us take an example of this type.

(iii) Frequency Curve

In simple words frequency curve is a smooth curve obtained by joining the points (not necessarily all points) of the frequency polygon such that

- Like frequency curve, it also starts from the baseline (horizontal axis) and ends at the baseline.
- Area under frequency curve remains approximately equal to the area under the frequency polygon.

In other words, let us try to explain the concept theoretically. Suppose we draw a sample of size n from a large population. Frequency curve is the graph of a continuous variable. So, theoretically, continuity of the variable implies that whatever small class interval we take, there will be some observations in that class interval. That is, in this case, there will be a large number of line segments and the frequency polygon tends to coincide with the smooth curve passing through these points as the sample size (n) increases. This smooth curve is known as frequency curve.

In the following example, we have drawn both frequency polygon and frequency curve to make the idea clear for you.

Example 4: Draw frequency polygon and frequency curve for the following frequency distribution.

Class Intervals	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
frequency	2	5	8	15	18	10	3	1

Solution: Frequency polygon and frequency curve for the above data is given below in Fig. 2.5.

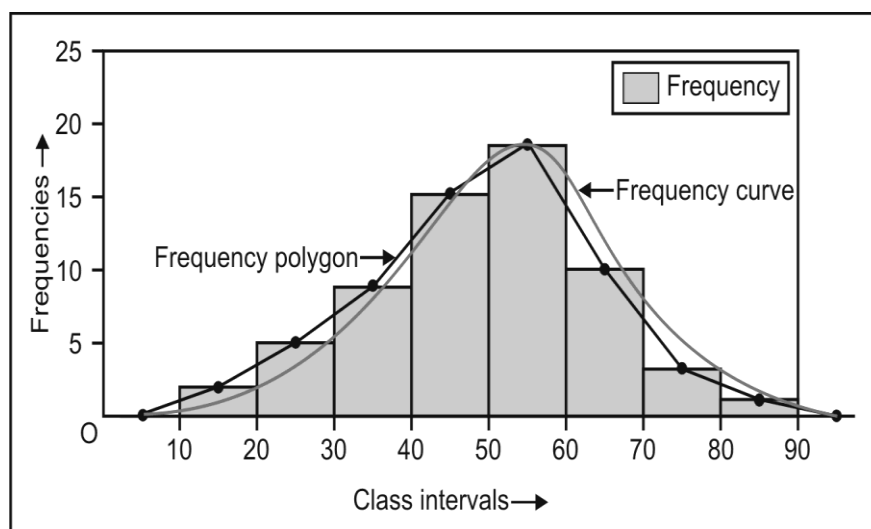


Fig. 2.5: Frequency Curve along with Frequency Polygon.

On the next page some important types of frequency curves are given which are generally obtained in the graphical presentations of frequency distributions. That is, symmetrical, positively skewed, negatively skewed, J shaped, U shaped, bimodal and multimodal frequency curves. You note that the shapes of these curves justify their names.

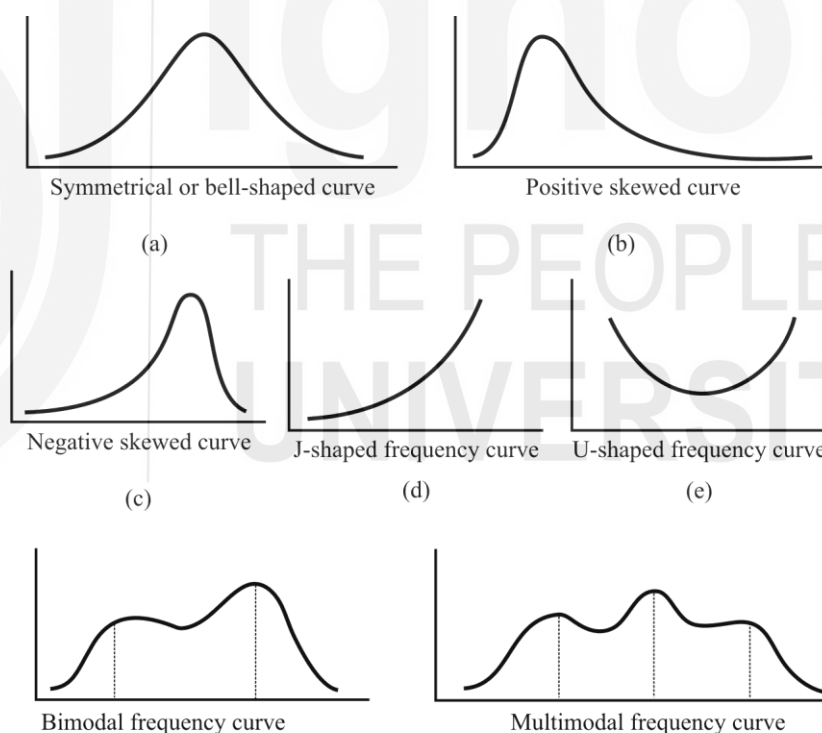


Fig. 2.6: Some Important Shapes of Curves

(iv) Cumulative Frequency Curves - Ogive

The concept of cumulative frequency distribution involves less than cumulative frequency distribution or more than cumulative frequency distribution. Here our aim is to study graphical presentation of less than and more than cumulative frequency distributions, which are known as less than frequency curve (or less than ogive) and more than frequency curve (or more than ogive) respectively.

For drawing less than cumulative frequency curve (or less than ogive), first of all the cumulative frequencies are plotted against the values (upper limits of the class intervals) up to which they correspond and then we simply join the points by line segments, curve thus obtained is known as less than ogive. Similarly, more than frequency curve (more than ogive) can be obtained by plotting more than cumulative frequencies against lower limits of the class intervals. As we have already mentioned within brackets that less than cumulative frequency curve and more than cumulative frequency curve are also called **less than ogive** and **more than ogive** respectively.

In other words we may define less than ogive and more than ogive as follow:

Less Than Ogive: If we plot the points with the upper limits of the classes as abscissae and the cumulative frequencies corresponding to the values less than the upper limits as ordinates and join the points so plotted by line segments, the curve thus obtained is nothing but known as “less than cumulative frequency curve” or “less than ogive”. It is a rising curve.

More Than Ogive: If we plot the points with the lower limits of the classes as abscissae and the cumulative frequencies corresponding to the values more than the lower limits as ordinates and join the points so plotted by line segments, the curve thus obtained is nothing but known as “more than cumulative frequency curve” or “more than ogive”. It is a falling curve.

Let us draw both the ogives (‘less than’ and ‘more than’) for the following frequency distribution of the weekly wages of the number of workers given in Table 2.7.

Table 2.7: A Sample Frequency Distribution

Weekly wages	0-10	10-20	20-30	30-40	40-50
No. of workers	45	55	70	40	10

Before drawing the ogives, we make a cumulative frequency distribution as given in Table 2.8

Table 2.8: Cumulative Frequency Distribution of Data of Table 2.7

Weekly wages	No. of workers	Less than Cumulative frequency distribution		More than Cumulative frequency distribution	
		Wages Less than	Number of workers	Wages More than	Number of workers
0-10	45	10	45	0	220
10-20	55	20	100	10	175
20-30	70	30	170	20	120
30-40	40	40	210	30	50
40-50	10	50	220	40	10

From the above data, we construct both the ogives as shown in Fig. 2.7 and Fig. 2.8. For “less than ogive” as shown on previous page in Fig. 2.7, we have plotted the points (10, 45), (20, 100), (30, 170), (40, 210), (50, 220) and then joined them by line segments. Similarly, for “more than ogive” as shown above in Fig. 2.8, we have plotted the points (0, 220), (10, 175), (20, 120), (30, 50), (40, 10), and then joined them by line segments.

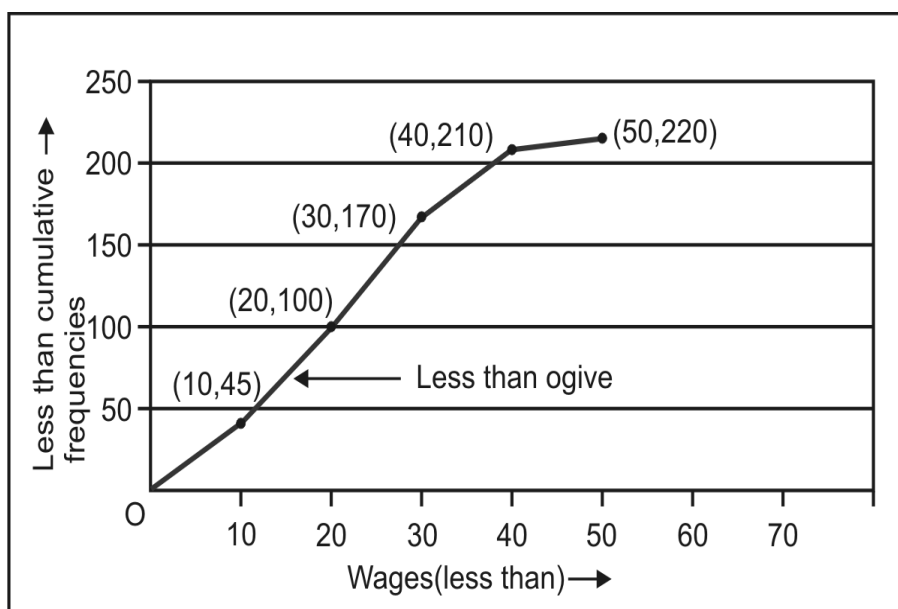


Fig. 2.7: Less Than Ogive.

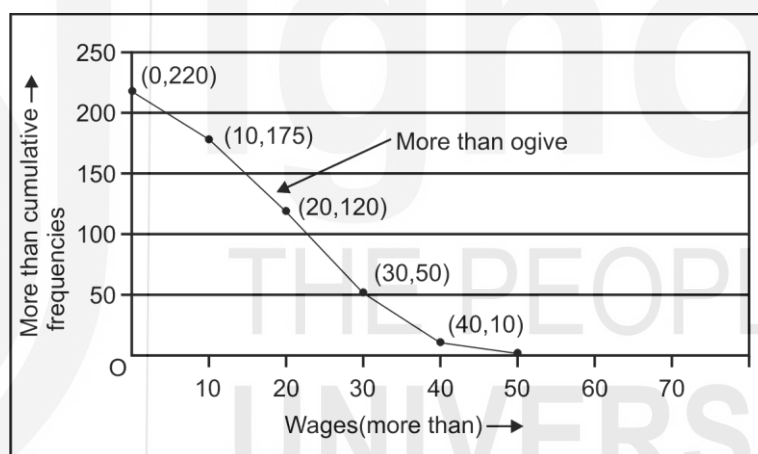


Fig. 2.8: More Than Ogive.

If we want to obtain a partition value, using ogives, we draw dotted horizontal line through that value at y-axis which corresponds to the partition value and then from the point, where it meets the less than ogive, we draw a dotted vertical line and let it meet the x-axis. The abscissa of the point, where it meets the x-axis is the required partition value. For example, suppose we want to find first quartile, then we draw a dotted horizontal line starting from y-axis at a point corresponding to $N/4$ and let it meet the “less than ogive”. From that point at “less than ogive”, we draw a dotted vertical line and let it meet the x-axis. The abscissa corresponding to this point is the first quartile. Similarly, for finding median or second quartile, we start drawing dotted horizontal line from y-axis at a point corresponding to $N/2$ and then we proceed as described above. Similarly, for third quartile $3N/4$ is taken in place of $N/2$. In this way, we may find any partition value.

Note 4: Median may also be obtained by drawing dotted vertical line through the point of intersection of both the ogives, when drawn on a single figure.

2.4.2 Stem-And-Leaf Plots

A stem-and-leaf display is very similar to a histogram but shows more information. The stem-and-leaf display summarises the shape of a set of data and provides the details regarding individual values. A stem-and-leaf display quickly summarises data while maintaining the individual data points. Now a day's use of stem-and-leaf displays is increasing, so let us formally define it in the next paragraph with some examples.

It has a vertical line of numbers obtained after removing the last digits (i.e. unit digits) from the given numbers called starting parts and for each starting part there is a horizontal line of numbers, i.e. the digits at the unit places of the given numbers called leaves. And each complete horizontal line including starting part and leaves is known as stem. The data displayed like this is nothing but known as stem-and-leaf display. The distance between the lowest values that are recorded in two consecutive stems is known as **stem width** or **category interval**, which plays very important role in stem-and-leaf displays.

Stem-and-leaf displays are used in many situations like series of scores of sports teams, series of temperature or rainfall over a period of time, series of classroom test scores, etc.

The following example will illustrate the above discussion more clearly.

Example 5: We have a set of values of the test scores of 22 students in a class as 11, 2, 28, 33, 48, 0, 42, 17, 24, 14, 0, 18, 26, 29, 35, 42, 22, 8, 28, 8, 46, 14. Draw a simple stem-and-leaf display by taking stem width 10.

Solution: Simple stem-and-leaf display for the given data can be shown as follows:

0		2 0 0 8 8	(5)
1		1 7 4 8 4	(5)
2		8 4 6 9 2 8	(6)
3		3 5	(2)
4		8 2 2 6	(4)

Total observations = 22

After arranging the leaves in ascending order of magnitude, we have

0		0 0 2 8 8
1		1 4 4 7 8
2		2 4 6 8 8 9
3		3 5
4		2 2 6 8

Here starting parts show the 'tens digits' and the leaves show the 'ones digits' in the above stem-and-leaf display. At a glance, one can see that 4 students got marks in the 40's in their test out of 50. Out of these four students two got 42 marks each, whereas the other two got 46 and 48 marks in the test. Fourth row (i.e. fourth stem) indicates that two students got marks in 30's in their test out of 50. And actual marks of these two students are 33 and 35. Similarly, we can get the information about the marks of the other students from successive rows (stems). When you count the total numbers of leaves, you may know how many students appeared in the test. The information is nicely organised when a stem-and-leaf display is used. Stem-and-leaf display provides a tool for specific

information in large sets of data, otherwise one would have a long list of marks to arrange and analyse.

Stem-and-Leaf Display for more than one Set of Data

Stem-and-leaf display is also used to compare two sets of data. That is known as 'back to back' stem-and-leaf display. For example, if you want to compare the batting scores of two cricket players, then stem-and-leaf display is right way to represent the data.

Example 6: Draw a stem-and-leaf display for batting scores of two players given below.

Player A	102, 61, 82, 88, 90, 63, 69, 85, 105, 93, 65, 94, 107, 97, 67
Player B	104, 62, 83, 95, 106, 95, 108, 63, 108, 82, 93, 109

Solution: The scores of two players can be compared with the help of back to back stem-and-leaf display as follows:

Leaf (Player A)	Starting part	Leaf (Player B)
1 3 5 7 9	6	2 3
2 5 8	8	2 3
0 3 4 7	9	3 5 5
2 5 7	10	4 6 8 8 9

Here column of starting parts is now in the middle and the leaves columns are to the right (player B) and left (player A) of the column of starting parts. You can see that the player B has more innings with a highest score than the player A. The player B has only 2 innings with scores of 62 and 63, while the player A has 5 innings with the scores of 61, 63, 65, 67 and 69. You can also see that player B has the highest score of 109, compared to player A with highest score of 107. Thus we see that presentation of the data by stem-and-leaf display provides us lot of information in very quick time.

Note: Stem-and-leaf display arranges the data in place values. They show the shape of data and can help to find median, mode and quartiles of data.

Check Your Progress 2

- Draw a histogram for the following frequency distribution
Wage (Rs): 0-5 5-10 10-15 15-20 20-25 25-30 30-35 35-40 40-45 45-50
No. of
Workers: 30 70 100 110 140 150 130 100 90 60
- Draw a frequency polygon from the following data
Wage: 0-5 5-10 10-15 15-20 20-25 25-30 30-35 35-40 40-45 45-50
No of
Workers: 20 50 80 110 140 150 120 150 100 80
- Draw a frequency curve from the following frequency distribution
Class Intervals: 1-2 2-3 3-4 4-5 5-6 6-7 7-8 8-9
Frequency: 3 9 20 25 18 12 6 3
- Draw two ogives from the following data
Class: 0-10 10-20 20-30 30-40 40-50 50-60 60-70 70-80 80-90
Frequency: 3 6 10 13 20 18 15 9
- Draw a stem-and-leaf display with the following marks obtained by 30 students.
77, 80, 82, 68, 65, 59, 61, 57, 50, 62, 61, 70, 69, 64, 67,
70, 62, 65, 65, 73, 76, 87, 80, 82, 83, 79, 79, 71, 80, 77

2.5 MEASURES OF CENTRAL TENDENCY

According to **Professor Bowley**, averages are “statistical constants which enable us to comprehend in a single effort the significance of the whole”. They throw light as to how the values are concentrated in the central part of the distribution. For this reason they are also called the measures of central tendency, an average is a single value which is considered as the most representative for a given set of data. Measures of central tendency show the tendency of some central value around which data tend to cluster.

Why use the Measure of Central Tendency?

The following are two main reasons for studying an average:

1. **To get a single representative:** A Measure of central tendency enables us to get a single value from the mass of data and also provides an idea about the entire data. For example, it is impossible to remember the height measurements of all students in a class. But if the average height is obtained, we get a single value that represents the entire class.
2. **To facilitate comparison:** Measures of central tendency enable us to compare two or more populations by reducing the mass of data in a single figure. The comparison can be made either at the same time or over a period of time. For example, if a subject has been taught in more than two classes, the average marks of those classes can be obtained, and a comparison can be made.

What should be the Properties of a Good Average?

The following are the properties of a good measure of average:

1. It should be simple to understand, as its purpose is to simplify complexity.
2. It should be easy to calculate so that it can be used as widely as possible.
3. It should be defined unambiguously so that if different people compute the average from same figures, they get the same answer.
4. It should be liable for algebraic manipulations. For example, if there are two sets of data and the individual information is available for both set, then one can be able to find the information regarding the combined set also then something is missing.
5. It should be least affected by sampling fluctuations. In other words, if we select 10 different groups of observations from same population and compute the average of each group, then we should expect to get approximately the same values. There may be little difference because of the sampling fluctuation only.
6. It should be based on all the observations, that is each and every observation is used for its calculation.
7. It should be possible to calculate even for open-end class intervals
8. It should not be affected by extremely small or extremely large observations. If one or two very small or very large observations affect the average i.e. either increase or decrease its value largely, then the average cannot be considered as a good average.

Different Measures of Central Tendency

The following are the various measures of central tendency:

1. Arithmetic Mean
2. Weighted Mean
3. Median
4. Mode
5. Geometric Mean
6. Harmonic Mean

The next sub-sections discuss them in more detail

2.5.1 Mean

Arithmetic mean (also called mean) is defined as the sum of all the observations divided by the number of observations. Arithmetic mean (AM) may be calculated for the following two types of data:

1. For Ungrouped Data

For ungrouped data, arithmetic mean may be computed by applying any of the following methods:

(1) Direct Method

Mathematically, if $x_1, x_2, \dots, x_n$ are the n observations then their mean is

$$\bar{X} = \frac{(x_1 + x_2 + x_3 + \dots + x_n)}{n}$$

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

If f_i is the frequency of x_i ($i=1, 2, \dots, k$), the formula for arithmetic mean would be

$$\bar{X} = \frac{(f_1 x_1 + f_2 x_2 + \dots + f_k x_k)}{(f_1 + f_2 + \dots + f_k)}$$

$$\bar{X} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i}$$

(2) Short-cut Method

The arithmetic mean can also be calculated by taking deviations from any arbitrary point “A”, in which the formula shall be

$$\bar{X} = A + \frac{\sum_{i=1}^n d_i}{n} \quad \text{where, } d_i = x_i - A$$

If f_i is the frequency of x_i ($i=1, 2, \dots, k$), the formula for arithmetic mean would be

$$\bar{X} = A + \frac{\sum_{i=1}^k f_i d_i}{\sum_{i=1}^k f_i}, \quad \text{where, } d_i = x_i - A$$

Here, k is the number of distinct observations in the distribution.

Note: Usually, the shortcut method is used when data values are large.

Example 7: Calculate mean of the weights of five students by using shortcut method.

54, 56, 70, 45, 50 (in kg)

Solution: For shortcut method, we use following formula

$$\bar{X} = A + \frac{\sum d_i}{n}, \quad \text{where } d_i = x_i - A$$

If 50 is taken as the assumed value A in the given data in then, for the calculation of d_i we prepare following table:

We have $A = 50$ then,

$$\begin{aligned} \bar{X} &= A + \frac{\sum_{i=1}^n d_i}{n} \\ &= 50 + \frac{25}{5} = 50 + 5 = 55 \end{aligned}$$

x	$d = x - A$
54	$54 - 50 = 4$
56	$56 - 50 = 6$
70	$70 - 50 = 20$
45	$45 - 50 = -5$
50	$50 - 50 = 0$
	$\sum_{i=1}^n d_i = 25$

2 For Grouped Data

Direct Method

If f_i is the frequency of x_i ($i = 1, 2, \dots, k$) where x_i is the mid value of the i^{th} class interval, the formula for arithmetic mean would be

$$\bar{X} = \frac{(f_1 x_1 + f_2 x_2 + \dots + f_k x_k)}{(f_1 + f_2 + \dots + f_k)}$$

$$\bar{X} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{N},$$

where, $N = f_1 + f_2 + \dots + f_k$

Short-cut Method

$$\bar{X} = A + \frac{\sum_{i=1}^k f_i d_i}{\sum_{i=1}^k f_i}, \quad \text{where, } d_i = x_i - A$$

Here, f_i would be representing the frequency of the i^{th} class, x_i is the mid-value of the i^{th} class and k is the number of classes.

Example 8: For the following data, calculate arithmetic mean using direct method:

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	3	5	7	9	4

Solution: We have the following distribution:

Class Interval	Mid Value x	Frequency f	fx
0-10	05	03	15
10-20	15	05	75
20-30	25	07	175
30-40	35	09	315
40-50	45	04	180
		$\sum_{i=1}^k f_i = N = 28$	$\sum_{i=1}^k f_i x_i = 760$

$$\text{Mean} = \frac{\sum_{i=1}^k f_i x_i}{N} = 760/28 = 27.143$$

Note on Arithmetic Mean

Arithmetic mean fulfills most of the properties of a good average except the last two. It is particularly useful when we are dealing with a sample as it is least affected by sampling fluctuations. It is the most popular average and should always be our first choice unless there is a strong reason for not using it.

2.5.2 Median

Median is that value of the variable which divides the whole distribution into two equal parts. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude. When the number of observations is odd then the median is the middle value of the data. For even number of observations, there will be two middle values. So we take the arithmetic mean of these two middle values. Number of the observations below and above the median, are same. Median is not affected by extremely large or extremely small values (as it corresponds to the middle value) and it is also not affected by open end class intervals. In such situations, it is preferable in comparison to mean. It is also useful when the distribution is skewed (asymmetric). Skewness will be discussed in the next unit.

1. Median for Ungrouped Data

Mathematically, if $x_1, x_2, \dots, x_n$ are the n observations then for obtaining the median first of all we have to arrange these n values either in ascending order or in descending order. When the observations are arranged in ascending or descending order, the middle value gives the median if n is odd. For even number of observations there will be two middle values. So we take the arithmetic mean of these two values.

$$M_d = \left(\frac{n+1}{2} \right)^{\text{th}} \text{ observation ; (when } n \text{ is odd)}$$

$$M_d = \frac{\left(\frac{n}{2}\right)^{\text{th}} \text{ observation} + \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ observation}}{2}; \text{ (when } n \text{ is even)}$$

For Ungrouped Data (when frequencies are given)

If x_i are the different value of variable with frequencies f_i then we calculate cumulative frequencies from f_i then median is defined by

$$M_d = \text{Value of variable corresponding to } \left(\frac{\sum f}{2}\right)^{\text{th}}$$

$$= \left(\frac{N}{2}\right)^{\text{th}} \text{ cumulative frequency.}$$

Note: If $N/2$ is not the exact cumulative frequency then value of the variable corresponding to next cumulative frequencies is the median.

Example 9: Find Median from the given frequency distribution

x	20	40	60	80
f	7	5	4	3

Solution: First we find cumulative frequency

x	f	c.f.
20	7	7
40	5	12
60	4	16
80	3	19
	$\sum_{i=1}^k f_i = 19$	

$$M_d = \text{Value of the variable corresponding to the } \left(\frac{19}{2}\right)^{\text{th}} \text{ cumulative frequency}$$

$$= \text{Value of the variable corresponding to 9.5 since 9.5 is not among c.f.}$$

So, the next cumulative frequency is 12 and the value of variable against 12 cumulative frequency is 40. So median is 40.

2. Median for Grouped Data

For class interval, first we find cumulative frequencies from the given frequencies and use the following formula for calculating the median:

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

where, L = lower class limit of the median class,

N = total frequency,

C = cumulative frequency of the pre-median class,

f = frequency of the median class, and

h = width of the median class.

Median class is the class in which the $(N/2)^{\text{th}}$ observation falls. If $N/2$ is not among any cumulative frequency then next class to the $N/2$ will be considered as median class.

Example 10: Calculate median for the data given in Example 8.

Solution: We first form a cumulative frequency table (cumulative frequency of a class gives the number of observations less than the upper limit of the class; strictly speaking, this is called cumulative frequency of less than type; we also have cumulative frequency of more than type which gives the number of observations greater than or equal to the lower limit of a class):

Class Interval	Frequency f	Cumulative Frequency (< type)
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28

$$\sum_{i=1}^k f_i = N = 28 \quad \Rightarrow \quad \frac{N}{2} = \frac{28}{2} = 14$$

Since 14 is not among the cumulative frequency so the class with next cumulative frequency i.e. 15, which is 20-30, is the median class.

We have

L = lower class limit of the median class = 20

N = total frequency = 28

C = cumulative frequency of the pre median class = 8

f = frequency of the median class = 7

h = width of median class = 10

Now substituting all these values in the formula of Median

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

$$M_d = 20 + \frac{14 - 8}{7} \times 10 = 28.57$$

Therefore, median is 28.57.

Note: Median is easy to understand and compute, and is not affected by extremely small or extremely large values; however, it does not utilise all the observations. For example, the median of 1, 2, 3 is 2. If observation 3 is replaced by any number higher than or equal to 2 and if the number 1 is replaced by any number lower than or equal to 2, the median value will be unaffected. This means 1 and 3 are not being utilised.

2.5.3 Mode

Highest frequent observation in the distribution is known as mode. In other words, mode is that observation in a distribution which has the maximum frequency. For example, when we say that the average size of shoes sold in a shop is 7 it is the modal size which is sold most frequently.

For Ungrouped Data

Mathematically, if $x_1, x_2, \dots, x_n$ are the n observations and if some of the observation are repeated in the data, say x_i is repeated highest times then we can say the x_i would be the mode value.

For Grouped Data:

Data where several classes are given, following formula of the mode is used

$$M_0 = L + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times h$$

where, L = lower class limit of the modal class,

f_1 = frequency of the modal class,

f_0 = frequency of the pre-modal class,

f_2 = frequency of the post-modal class, and

h = width of the modal class.

Modal class is that class which has the maximum frequency.

Example 11: For the data given in Example 8, calculate mode.

Solution: Here the frequency distribution is

Class Interval	Frequency
0-10	3
10-20	5
30-40	7
40-50	9
50-60	4

Corresponding to highest frequency 9 modal class is 40-50 and we have

$$L = 40, f_1 = 9, f_0 = 7, f_2 = 4, h = 10$$

Applying the formula,

$$\begin{aligned} \text{Mode} &= 40 + \frac{(9 - 7)}{(2 \times 9 - 7 - 4)} \times 10 \\ &= 42.86 \end{aligned}$$

Note: Mode is the easiest average to understand and also easy to calculate and is not affected by extreme values. Mode can be calculated even if the classes are of unequal width. However, a distribution can have more than one mode, and it does not utilize all the observations. Further, Mode is not amenable to algebraic treatment; an is greatly affected by sampling fluctuations.

2.5.4 Relationship between Mean, Median and Mode

For a symmetrical distribution the mean, median and mode coincide. But if the distribution is moderately asymmetrical, there is an empirical relationship between them. The relationship is

$$\text{Mean} - \text{Mode} = 3 (\text{Mean} - \text{Median})$$

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Note: Using this formula, we can calculate mean/median/mode if other two of them are known.

2.5.5 Geometric and Harmonic Mean

Geometric Mean: The geometric mean (GM) of n observations is defined as the n -th root of the product of the n observations. It is useful for averaging ratios or proportions. It is the ideal average for calculating index numbers (index numbers are economic barometers which reflect the change in prices or commodity consumption in the current period with respect to some base period taken as standard). It fails to give the correct average if an observation is zero or negative.

1. For Ungrouped Data

If $x_1, x_2, \dots, x_n$ are the n observations of a variable X then their geometric mean is

$$GM = \sqrt[n]{x_1 x_2 \dots x_n}$$

$$GM = (x_1 x_2 \dots x_n)^{\frac{1}{n}}$$

Taking log of both sides

$$\log GM = \frac{1}{n} \log (x_1 x_2 \dots x_n)$$

$$\log GM = \frac{1}{n} (\log x_1 + \log x_2 + \dots + \log x_n)$$

$$\Rightarrow GM = \text{Antilog} \left(\frac{1}{n} \sum_{i=1}^n \log x_i \right)$$

Example 12: Find geometric mean of 2, 4, 8.

Solution: $GM = (x_1 x_2 \dots x_n)^{\frac{1}{n}}$

$$GM = (2 \times 4 \times 8)^{\frac{1}{3}} = (64)^{\frac{1}{3}} = 4$$

Thus, the geometric mean is 4.

2. For Grouped data

If $x_1, x_2, \dots, x_k$ are k values (or mid values in case of class intervals) of a variable X with frequencies $f_1, f_2, \dots, f_k$ then

$$GM = (x_1^{f_1} x_2^{f_2} \dots x_k^{f_k})^{\frac{1}{\sum f_i}}$$

$$GM = (x_1^{f_1} x_2^{f_2} \dots x_k^{f_k})^{\frac{1}{N}}$$

where $N = f_1 + f_2 + \dots + f_k$

Taking log of both sides

$$\log GM = \frac{1}{N} \log (x_1^{f_1} x_2^{f_2} \dots x_k^{f_k})$$

$$\log GM = \frac{1}{N} (f_1 \log x_1 + f_2 \log x_2 + \dots + f_k \log x_k)$$

$$\Rightarrow GM = \text{Antilog} \left(\frac{1}{N} \sum_{i=1}^k f_i \log x_i \right)$$

Example 13: For the data in Example 8, calculate geometric mean.

Solution:

Class	Mid Value (x)	Frequency (f)	log x	f × log x
0-10	5	3	0.6990	2.0970
10-20	15	5	1.1761	5.8805
20-30	25	7	1.3979	9.7853
30-40	35	9	1.5441	13.8969
40-50	45	4	1.6532	6.6128
		N = 28		$\sum f_i \log x_i = 38.2725$

Using the formula

$$\begin{aligned}
 \text{GM} &= \text{Antilog} \left(\frac{1}{N} \sum_{i=1}^k f_i \log x_i \right) \\
 &= \text{Antilog} \left(\frac{38.2725}{28} \right) \\
 &= \text{Antilog} (1.3669) \\
 &= 23.28
 \end{aligned}$$

Note: Geometric mean uses all the observations and gives more weight to small items. It is not affected greatly by sampling fluctuations. However, it is difficult to calculate; and becomes imaginary for an odd number of negative observations and becomes zero or undefined if a single observation is zero.

Harmonic Mean: The harmonic mean (HM) is defined as the reciprocal (inverse) of the arithmetic mean of the reciprocals of the observations of a set.

1. For Ungrouped Data

If $x_1, x_2, \dots, x_n$ are the n observations of a variable X , then their harmonic mean is

$$\begin{aligned}
 \text{HM} &= \frac{1}{\frac{1}{n} \left[\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right]} \\
 \text{HM} &= \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}
 \end{aligned}$$

Example 14: Calculate the Harmonic mean of 1, 3, 5, 7

Solution: Formula for harmonic mean is

$$\text{HM} = \frac{1}{\frac{1}{n} \left[\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right]}$$

$$\begin{aligned}
 HM &= \frac{1}{\frac{1}{4} \left[\frac{1}{1} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} \right]} \\
 &= 4 / (1.0000 + 0.3333 + 0.2000 + 0.1428) \\
 &= 4/1.6761 = 2.39
 \end{aligned}$$

2. For Grouped Data

If $x_1, x_2, \dots, x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, \dots, f_k$, then

$$\begin{aligned}
 HM &= \frac{1}{\frac{1}{N} \left[\frac{f_1}{x_1} + \frac{f_2}{x_2} + \dots + \frac{f_k}{x_k} \right]} \\
 HM &= \frac{N}{\sum_{i=1}^k \frac{f_i}{x_i}} \quad \text{where, } N = \sum_{i=1}^k f_i
 \end{aligned}$$

When equal distances are travelled at different speeds, the average speed is calculated by the harmonic mean. It cannot be calculated if an observation is zero.

Example 15: For the data given in Example 8, calculate the harmonic mean.

Solution:

Class	Mid Value (x)	Frequency (f)	f/x
0-10	5	3	0.600
10-20	15	5	0.330
20-30	25	7	0.280
30-40	35	9	0.257
40-50	45	4	0.088
		$N = \sum f = 28$	$\sum f/x = 1.555$

Using the formula,

$$\begin{aligned}
 HM &= \frac{N}{\sum_{i=1}^k \frac{f_i}{x_i}} \\
 &= 28/1.555 = 17.956
 \end{aligned}$$

Note: Harmonic mean also uses all the observations, and gives greater importance to small items; however, it is difficult to understand and compute.

Relations between AM, GM and HM

We state the relationships without any proof. You may refer to the references for proof.

Relation 1: $AM \geq GM \geq HM$

Relation 2: $GM = \sqrt{AM \cdot HM}$

Check Your Progress 3

- 1 Find arithmetic mean of the distribution of marks given below:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	6	9	17	10	8

2. Find the missing frequency when median is given as Rs 50.

Expenditure (Rs)	0-20	20-40	40-60	60-80	80-100
No. of families	5	15	30	--	8

- 3 In an asymmetrical distribution the mode and mean are 35.4 and 38.6 respectively. Calculate the median.

.....

2.6 PARTITION VALUES

Partition values are those values of variable which divide the distribution into a certain number of equal parts. Here it may be noted that the data should be arranged in ascending or descending order of magnitude. Commonly used partition values are quartiles, deciles and percentiles. For example, quartiles divide the data into four equal parts. Similarly, deciles and percentiles divide the distribution into ten and hundred equal parts, respectively.

2.6.1 Quartiles

Quartiles divide whole distribution into four equal parts. There are three quartiles- 1st Quartile denoted as Q_1 , 2nd Quartile denoted as Q_2 and 3rd Quartile as Q_3 , which divide the whole data in four parts. 1st Quartile contains the $\frac{1}{4}$ part of data, 2nd Quartile contains $\frac{1}{2}$ of the data and 3rd Quartile contains the $\frac{3}{4}$ part of data. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude.

For Ungrouped Data

For obtaining the quartiles first of all we have to arrange the data either in ascending order or in descending order of their magnitude. Then, we find the $(N/4)^{th}$, $(N/2)^{th}$ and $(3N/4)^{th}$ placed item in the arranged data for finding out the 1st Quartile (Q_1), 2nd Quartile (Q_2) and 3rd Quartile (Q_3) respectively. The value of the $(N/4)^{th}$ placed item would be the 1st Quartile (Q_1), value of $(N/2)^{th}$ placed item would be the 2nd Quartile (Q_2) and value of $(3N/4)^{th}$ placed item would be the 3rd Quartile (Q_3) of that data.

Mathematically, if $x_1, x_2, \dots, x_N$ are N values of a variable X then 1st Quartile (Q_1), 2nd Quartile (Q_2) and 3rd Quartile (Q_3) are defined as:

First Quartile (Q_1) = $\frac{N}{4}$ th placed item in the arranged data

Second Quartile (Q_2) = $\frac{N}{2}$ th placed item in the arranged data

Third Quartile (Q_3) = $3 \left(\frac{N}{4} \right)$ th placed item in the arranged data

For Grouped Data

If $x_1, x_2, \dots, x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, \dots, f_k$, then first of all we form a cumulative frequency distribution. After that we determine the i^{th} quartile class as similar as we do in case of median.

The i^{th} quartile is denoted by Q_i and defined as

$$Q_i = L + \frac{\left(\frac{iN}{4} - C \right)}{f} \times h \quad \text{for } i = 1, 2, 3$$

where, L = lower class limit of i^{th} quartile class,

h = width of the i^{th} quartile class,

N = total frequency,

C = cumulative frequency of pre i^{th} quartile class, and

f = frequencies of i^{th} quartile class.

“ i ” denotes i^{th} quartile class. It is the class in which $\left(\frac{i \times N}{4} \right)^{\text{th}}$ observation falls in cumulative frequency. It is easy to see that the second quartile ($i = 2$) is the median.

2.6.2 Deciles

Deciles divide whole distribution in to ten equal parts. There are nine deciles. $D_1, D_2, \dots, D_9$ are known as 1st Decile, 2nd Decile, ..., 9th Decile respectively and i^{th} Decile contains the $(iN/10)^{\text{th}}$ part of data. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude.

For Ungrouped Data

For obtaining the deciles first of all we have to arrange the data either in ascending order or in descending order of their magnitude. Then, we find the $(1N/10)^{\text{th}}, (2N/10)^{\text{th}}, \dots, (9N/10)^{\text{th}}$ placed item in the arranged data for finding out the 1st decile (D_1), 2nd decile (D_2), ..., 9th decile (D_9) respectively. The value of the $(N/10)^{\text{th}}$ placed item would be the 1st decile, value of $(2N/10)^{\text{th}}$ placed item would be the 2nd decile. Similarly, the value of $(9N/10)^{\text{th}}$ placed item would be the 9th decile of that data.

Mathematically, if $x_1, x_2, \dots, x_N$ are N values of a variable X then the i^{th} decile is defined as:

$$i^{\text{th}} \text{ Decile } (D_i) = \frac{iN}{10} \text{th placed item in the arranged data } (i = 1, 2, 3, \dots, 9)$$

For Grouped Data

If $x_1, x_2, \dots, x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, \dots, f_k$, then first of all we form a cumulative frequency distribution. After that we determine the i^{th} deciles classes as similar as we do in case of quartiles.

The i^{th} decile is denoted by D_i and given by

$$D_i = L + \frac{\left(\frac{iN}{10} - C\right)}{f} \times h \quad \text{for } i = 1, 2, \dots, 9$$

where, L = lower class limit of i^{th} decile class,

h = width of the i^{th} decile class,

N = total frequency,

C = cumulative frequency of pre i^{th} decile class; and

f = frequency of i^{th} decile class.

“ i ” denotes i^{th} decile class. It is the class in which $\left(\frac{i \times N}{10}\right)^{\text{th}}$ observation falls in cumulative frequency. It is easy to see that the fifth quartile ($i=5$) is the median.

2.6.3 Percentiles

Percentiles divide whole distribution into 100 equal parts. There are ninety nine percentiles. $P_1, P_2, \dots, P_{99}$ are known as 1st percentile, 2nd percentile, ..., 99th percentile and i^{th} percentile contains the $(iN/100)^{\text{th}}$ part of data. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude.

For Ungrouped Data

For obtaining the percentiles first of all we have to arrange the data either in ascending order or in descending order of their magnitude. Then, we find the $(1N/100)^{\text{th}}, (2N/100)^{\text{th}}, \dots, (99N/100)^{\text{th}}$ placed item in the arranged data for finding out the 1st percentile (P_1), 2nd percentile (P_2), ..., 99th percentile (P_{99}) respectively. The value of the $(N/100)^{\text{th}}$ placed item would be the 1st percentile, value of $(2N/100)^{\text{th}}$ placed item would be the 2nd percentile. Similarly, the value of $(99N/100)^{\text{th}}$ placed item would be the 99th percentile of that data.

Mathematically, if $x_1, x_2, \dots, x_N$ are N values of a variable X then the i^{th} percentile is defined as:

$$i^{\text{th}} \text{ Percentile } (P_i) = \frac{iN}{100} \text{th placed item in the arranged data } (i = 1, 2, \dots, 99)$$

For Grouped Data

If $x_1, x_2, \dots, x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, \dots, f_k$, then first of all we form a cumulative frequency distribution. After that we determine the i^{th} percentile class as similar as we do in case of median. The i^{th} percentile is denoted by P_i and given by

$$P_i = L + \frac{\left(\frac{iN}{100} - C\right)}{f} \times h \quad \text{for } i = 1, 2, \dots, 99$$

where, L = lower limit of i^{th} percentile class,

h = width of the i^{th} percentile class,

N = total frequency,

C = cumulative frequency of pre i^{th} percentile class; and

f = frequency of i^{th} percentile class.

“ i ” denotes i^{th} percentile class. It is the class in which $\left(\frac{i \times N}{100}\right)^{\text{th}}$ observation falls in cumulative frequency. It is easy to see that the fiftieth percentile ($i = 50$) is the median.

Example 16: For the data given in Example 8, calculate the first and third quartiles.

Solution: First we find cumulative frequency given in the following cumulative frequency table:

Class Interval	Frequency	Cumulative Frequency (< type)
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28
	$N = \sum f = 28$	

Here, $N/4 = 28/4 = 7$. The 7th observation falls in the class 10-20. So, this is the first quartile class. $3N/4 = 21^{\text{th}}$ observation falls in class 30-40, so it is the third quartile class.

For first quartile $L = 10$, $f = 5$, $C = 3$, $N = 28$

$$Q_1 = 10 + \frac{(7-3)}{5} \times 10 = 18$$

For third quartile $L = 30$, $f = 9$, $C = 15$

$$Q_3 = 30 + \frac{(21-15)}{9} \times 10 = 36.67$$

2.6.4 Boxplots

In descriptive statistics, a box plot also known as a box-and-whisker plot, is a convenient way of graphically representing numerical data. It represents the data through five-number summary, i.e. the smallest observation (sample minimum), lower quartile (Q_1), median (Q_2), upper quartile (Q_3), and the largest observation (sample maximum). Box plots are used to describe a distribution generally when it is extremely skewed or multimodal. A box plot also indicates which observation(s), if any, might be considered as outliers. A box plot is a quick graphic approach for examining one or more sets of data.

Box plots display differences between populations without making any assumptions of the distributions. The spacing between the different parts of the box helps in indicating the degree of dispersion (spread) and skewness in the data, and identifies outliers. Box plots can be drawn either horizontally or vertically. Here we will draw box plots vertically.

Method of Construction of the Box Plots

Box plots are useful for identifying key values while comparing two or more distributions. To understand more clearly the method of constructing the box plots, let us consider the following data of 37 students in a class who were examined by a game with a box containing some balls. Their task was to select a ball from one box placed at one corner and put it in another blank box placed at another corner of the class as quickly as possible, and their times (in seconds) were recorded. The scores were compared for 16 boys and 21 girls who participated in a game. Observed data are given in the following table:

Time (in seconds) for completing the given task

Boys	18, 19, 20, 22, 24, 25, 26, 16, 17, 19, 25, 27, 28, 23, 23, 31
Girls	15, 17, 18, 19, 20, 21, 23, 14, 17, 18, 19, 20, 21, 24, 19, 16, 17, 18, 20, 22, 28

The method of construction of separate box plots for the data of boys and girls is discussed below:

There are several ways of constructing a box plot. The first relies on the quartiles, lowest and greatest values in the distribution of scores. Fig. 2.9 shows how these three statistics are used for the above example. We draw a box plot for each gender extending from the 1st quartile to the 3rd quartile. The 2nd quartile is drawn inside the box. Therefore,

- The bottom of each box is the 1st quartile,
- The top is the 3rd quartile,
- The line in the middle is the 2nd quartile.
- A line extending from the point corresponding to the smallest observation to 1st quartile is drawn and known as lower whisker.
- A line extending from the 3rd quartile to the point corresponding to the largest observation is drawn and is known as upper whisker.

Let us arrange the given data for each of the gender in ascending order as shown in following table to find out the above components, i.e. Q_1, Q_2, Q_3 , smallest observation and largest observation.

Gender	Times (in Seconds)
Boys	16, 17, 18, 19, 19, 20, 22, 22, 23, 23, 24, 25, 25, 27, 28, 31
Girls	14, 15, 16, 17, 17, 17, 18, 18, 18, 19, 19, 19, 20, 20, 20, 21, 21, 22, 23, 24, 28

For Boys

The lowest or smallest observation = $x_s = 16$.

$$\begin{aligned}\text{First quartile } (Q_1) &= \left(\frac{16+1}{4} \right)^{\text{th}} \text{ item} \\ &= 4.25^{\text{th}} \text{ item} = 4^{\text{th}} \text{ item} + 0.25 (5^{\text{th}} \text{ item} - 4^{\text{th}} \text{ item}) \\ &= 19 + 0.25 (19 - 19) = 19\end{aligned}$$

$$\begin{aligned}\text{Second quartile} = (Q_2) &= 2 \left(\frac{16+1}{4} \right)^{\text{th}} \text{ item} \\ &= 8.5^{\text{th}} \text{ item} = \text{Mean of } 8^{\text{th}} \text{ and } 9^{\text{th}} \text{ items} \\ &= \frac{22 + 23}{2} = 22.5\end{aligned}$$

$$\text{Third quartile } (Q_3) = 3 \left(\frac{16+1}{4} \right)^{\text{th}} \text{ item}$$

$$= 12.75^{\text{th}} \text{ item} = 12^{\text{th}} \text{ item} + 0.75 (13^{\text{th}} \text{ item} - 12^{\text{th}} \text{ item})$$

$$= 25 + 0.75 (25 - 25) = 25$$

The largest observation = $x_l = 31$

For Girls

The lowest or smallest observation = $x_s = 14$

$$\text{First quartile } (Q_1) = \left(\frac{21+1}{4} \right)^{\text{th}} \text{ item}$$

$$= 5.5^{\text{th}} \text{ item} = \text{mean of } 5^{\text{th}} \text{ and } 6^{\text{th}} \text{ items} = \frac{17+17}{2} = 17$$

$$\text{Second quartile } (Q_2) = 2 \left(\frac{21+1}{4} \right)^{\text{th}} \text{ item} = 11^{\text{th}} \text{ item} = 19$$

$$\text{Third quartile } (Q_3) = 3 \left(\frac{21+1}{4} \right)^{\text{th}} \text{ item}$$

$$= 16.5^{\text{th}} \text{ item} = \text{mean of } 16^{\text{th}} \text{ and } 17^{\text{th}} \text{ item} = \frac{21+21}{2} = 21$$

The largest observation = $x_l = 28$

Box plots for boys and girls on the basis of the above findings are shown below in Fig. 2.9.

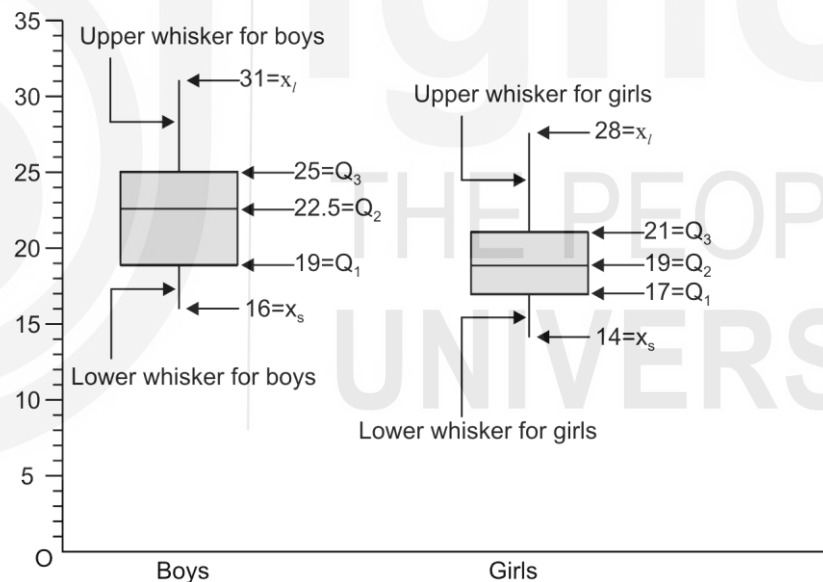


Fig. 2.9: The Box Plots with the Whiskers for Boys and Girls

Components of a Box Plot

Let us now discuss the various components and terminologies which are used in drawing the various types of box plots.

1. **Upper and Lower Hinges:** The upper and lower hinges are constructed to represent the third quartile and first quartile, respectively. For the example discussed above, the values of upper and lower hinges are 21 and 17 for the girls, whereas for the boys, they are 25 and 19, respectively.
2. **H-Spread:** This is calculated by taking the difference between the upper and lower hinges. The H-Spread shows the spread of the elements of the data between the first quartile and the third quartile. For the data related to boys

in the example discussed above, the H-spread is $25 - 19 = 6$, whereas for girls it is $21 - 17 = 4$.

3. **Whiskers:** The lines extending above and below the box are called whiskers. Lower whisker extends from the point corresponding to the smallest observation to Q_1 and the upper whisker extends from Q_3 to the largest observation.
4. **Step:** The step is calculated by multiplying the H-spread by 1.5. For the data of boys in the above example, the value of 1 step is $1.5 \times 6 = 9$, whereas for girls, it is $1.5 \times 4 = 6$.
5. **Upper and Lower Inner Fences:** The upper and lower inner fences are calculated by adding one step to the upper hinge and subtracting 1 step from the lower hinge, respectively. In other words, upper inner fence is equal to the 1 step ahead from upper hinge whereas the lower inner fence is one step before the lower hinge. For the data of boys in the example discussed above, the upper and lower inner fences are $25 + 9 = 34$ and $19 - 9 = 10$, whereas for the data of girls the upper and lower inner fences are $21 + 6 = 27$ and $17 - 6 = 11$, respectively.
6. **Upper and Lower Outer Fences:** The upper and lower outer fences are calculated by adding two steps to the upper hinge and subtracting 2 steps from the lower hinge, respectively. For the data of boys in the example discussed above the upper and lower outer fence are $25 + 2 \times 9 = 43$ and $19 - 2 \times 9 = 1$, whereas for the data of girls they are, $21 + 2 \times 6 = 33$ and $17 - 2 \times 6 = 5$, respectively.
7. **Upper and Lower Adjacent:** The upper and lower adjacent are used to represent the largest and the smallest observations of the data. For the data of boys in the example discussed above upper and lower adjacent are 31 and 16 whereas for the girls data they are 28 and 14, respectively.

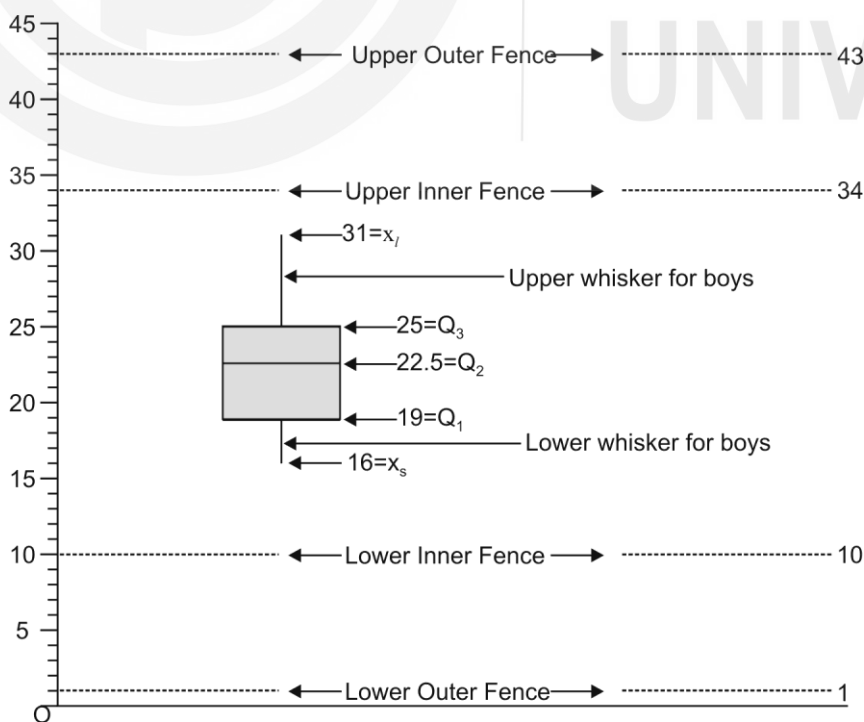


Fig. 2.10 Box Plot for Boys with Inner and Outer Fences

8. **Outside Value:** The outside value is a value which is beyond an inner fence but not beyond an outer fence. This is used to represent the scattered values of the data. To represent these values circles are used.
9. **Far Out Value:** A value which is beyond the upper outer fence or lower outer fence is known as far out value. To represent these values, the asterisks are used.

Box plot with the components discussed above for the data of boys is shown in Fig. 2.10.

Box plot with the components discussed above for the data of girls is shown in Fig. 2.11.

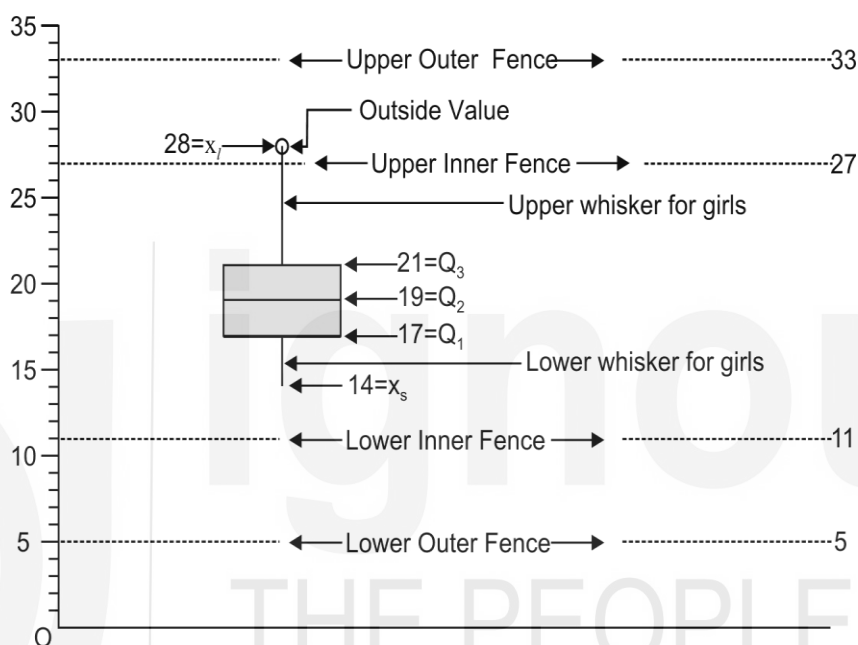


Fig. 2.11 Box Plot for Girls with Inner and Outer Fences

Box Plots with Outliers

Outside value(s) and far out value(s) are known as extreme observations. If the extreme observations are present in the data then those observations may be represented in box plot by an individual mark. The extreme observations are described as outliers when they are represented in box plots. The outside value is a value which is beyond an inner fence but not beyond an outer fence whereas the far out value is a value which is beyond the lower and upper outer fences. The individual marks for extreme values can be plotted above and below to the whiskers in box plot. Specially, outside values are indicated by small circles. The box plot for girls data indicating the outside value by a circle is shown in Fig. 2.12.

Box Plots with + Signs

One more component which is to be included in box plots are the value of some important parameters like, mean, mode, etc. We indicate the mean score for a group of values by inserting a plus sign in box plots. For the example mean in case of boys data is 20.875 and mean in case of girls data is 19.33, which are shown by a plus sign in Fig. 2.13.

Fig. 2.13 provides a revealing summary of the data. Since half of the scores are between the hinges (recall that the hinges are the first and third quartiles), we

see that half of the girl's times are between 17 and 21, whereas half of the boy's times are between 19 and 25.

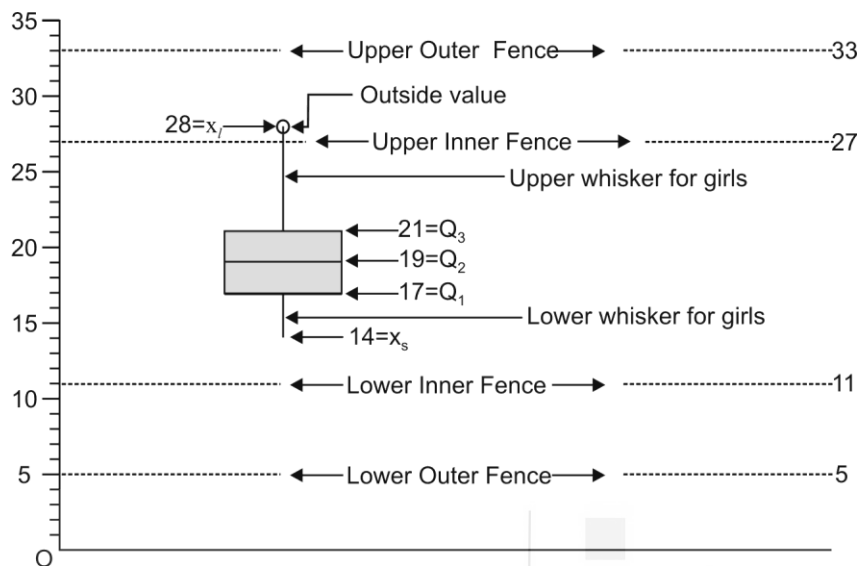


Fig. 2.12: The Box Blot for Girls with the Outlier.

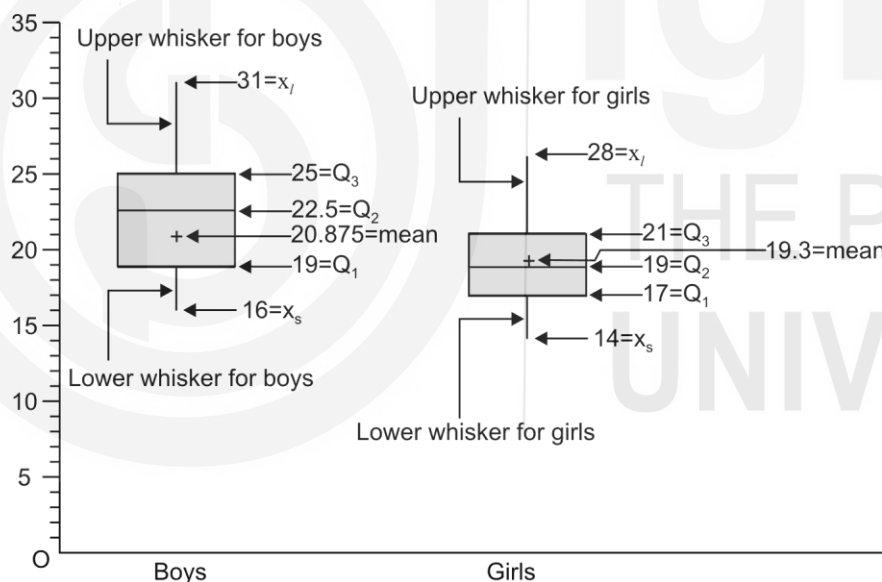


Fig. 2.13: The Box Plot with Whisker and + Signs for Boys and Girls Data.

Check Your Progress 4

- 1 Draw a box plot for the given data:
17, 15, 17, 20, 13, 15, 15, 16, 16, 15, 19, 12, 19, 14, 11, 14, 16, 10, 19, 18,
20, 14, 17, 19, 16, 22, 21, 23, 11, 12, 18, 13, 12, 25, 14, 15, 25, 17, 10, 21
- 2 Draw a box plot for the given data:
31, 42, 22, 27, 33, 27, 37, 28, 34, 44, 25, 39, 26, 31, 26, 33, 46, 48, 50

2.7 SUMMARY

In this unit, we have discussed:

1. How to make frequency distribution for different types of data.
 2. How to create class intervals for the given data.
 3. How to graphically represent frequency distribution of data.
 4. The properties of a good measures of central tendencies.
 5. How to compute different measures of central tendency.
 6. The different kinds of partition values and use of boxplots.
-

2.8 SOLUTIONS / ANSWERS

Check Your Progress 1

- 1 Let us determine the suitable class interval with the help of the following formula:

$$i = \frac{\text{Range}}{1 + 3.322 \log N}$$

$$\text{Range} = 59 - 06 = 53, N = 30$$

$$i = \frac{53}{1 + 3.322 \log 30} = \frac{53}{1 + 4.91} = 8.97 \approx 9$$

Since values like 3, 7, 9 etc., should be avoided and therefore, we will take 10 as the class interval and hence let us take the first class as 5-15 and thus the following table is formed:

Continuous Frequency Distribution of 30 Students

Heights (cm)	Tally Mark	Frequency
05-15		5
15-25		2
25-35		8
35-45		5
45-55		5
55-65		5
Total		30

- 2 As the least value is 3 and the highest value is 93, so using

$$i = \frac{\text{Range}}{1 + 3.322 \log N} = \frac{93-3}{1 + 3.322 \log 60} \approx 13.03 \approx 13$$

Since, values like 3, 7, 9, 11, 13 etc., should be avoided and therefore, we will take 14 as class interval and hence let us take the first class as 0-14 and thus the following table is formed.

Continuous Frequency Distribution of 60 Students According to their Heights

Profit (Cr)	Tally Mark	Frequency
0-14		06
15-29		04
30-44		16
45-59		10
60-74		14
75-89		07
90-104		03
Total		60

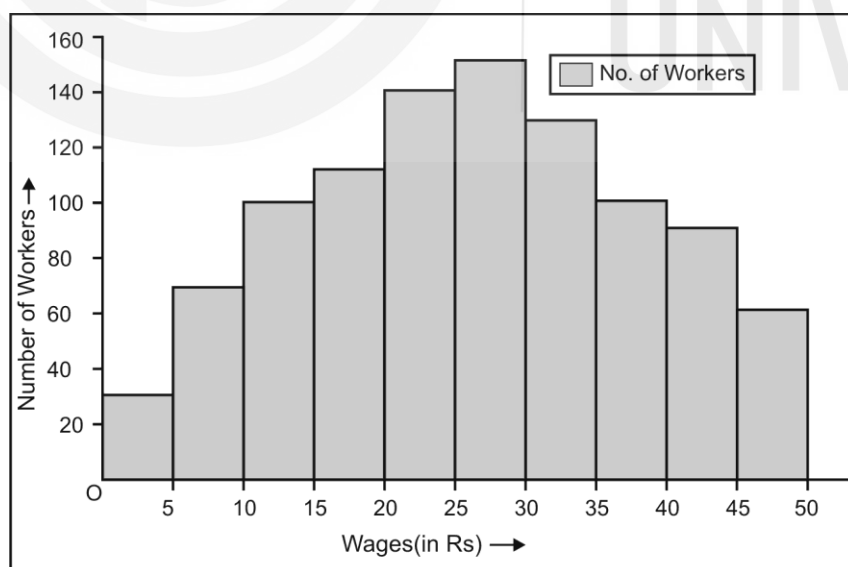
- 3 The following table illustrates the way of classification of data according to the exclusive method and principle of correction factor in classification.

Continuous Frequency Distribution of 60 Students According to their Heights

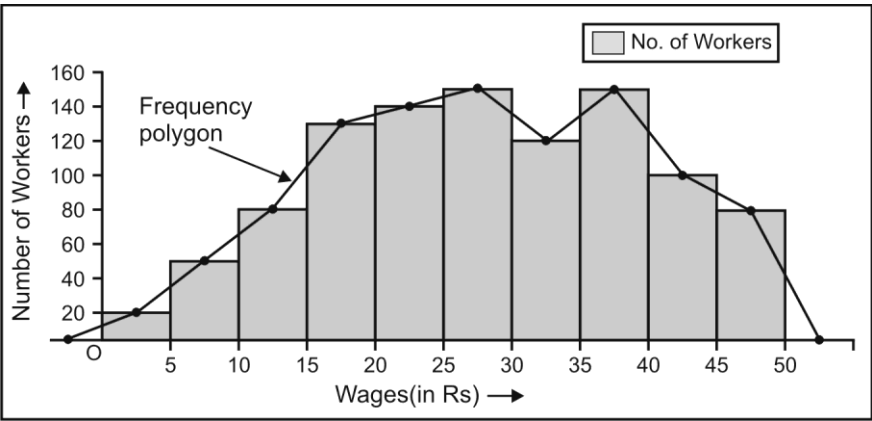
Profit (Cr)	Tally Mark	Frequency
0-14.5		06
14.5-29.5		04
29.5-44.5		16
44.5-59.5		10
59.5-74.5		14
74.5- 89.5		07
89.5-104.5		03
Total		60

Check Your Progress 2

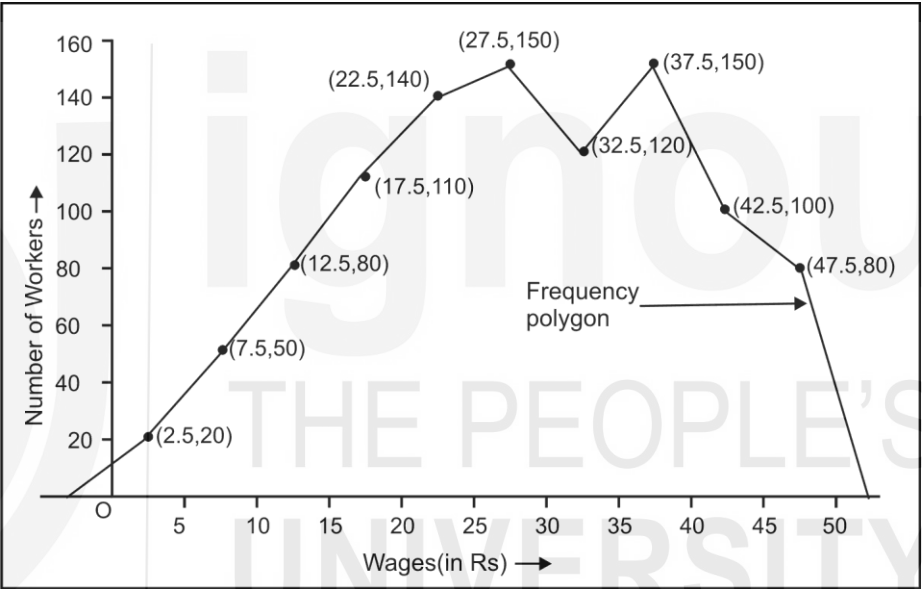
- 1 Histogram of the given data is given below:



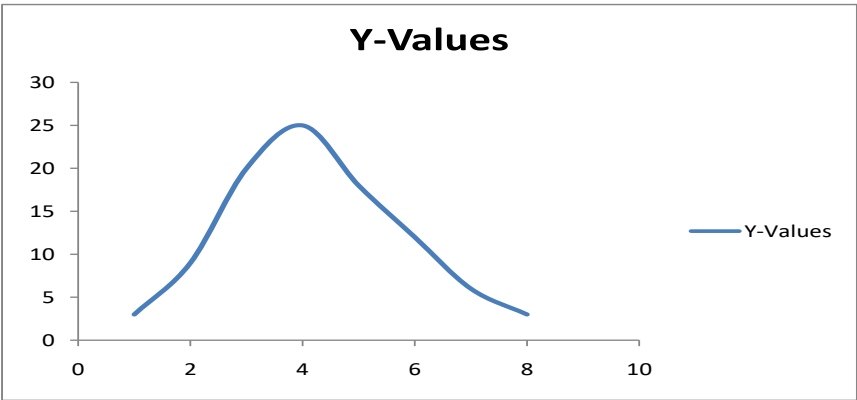
2 Frequency polygon of the given data by first drawing histogram is given below:



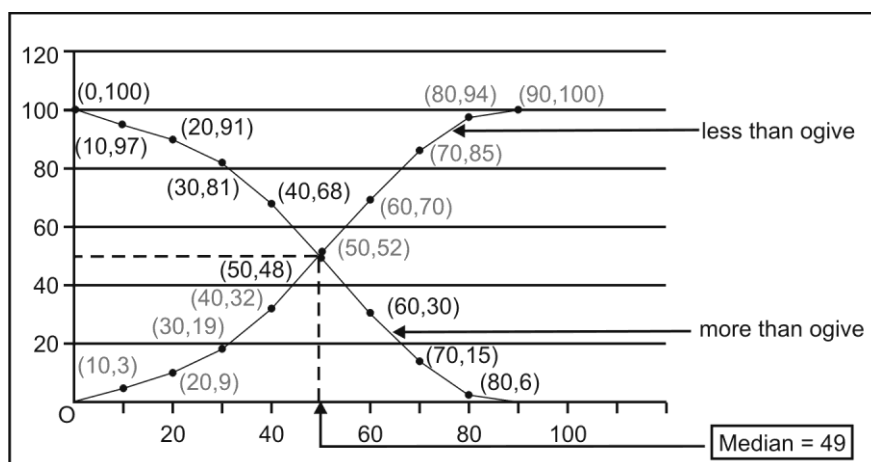
Frequency polygon can also be drawn without histogram as given below:



3 Frequency curve for the given data is given below.



4 Two ogives for the given data are given below:



5. Stem-and-leaf display of the given data of marks obtained by 30 students is given below.

```

5 | 9 7 0
6 | 8 5 1 2 1 9 4 7 2 5 5
7 | 7 0 0 3 6 9 9 1 7
8 | 0 2 7 0 2 3 0
  
```

After arranging the leaves in ascending order of magnitude, we have

```

5 | 0 7 9
6 | 1 1 2 2 4 5 5 5 7 8 9
7 | 0 0 1 3 6 7 7 9 9
8 | 0 0 0 2 2 3 7
  
```

Check Your Progress 3

1 We have the following frequency distribution:

Marks	No. of Students (f)	Mid Points x	fx
0-10	6	5	30
10-20	9	15	135
20-30	17	25	425
30-40	10	35	350
40-50	8	45	360
	$\sum f_i = 50$		$\sum f_i x_i = 1300$

$$\bar{X} = \frac{\sum_{i=1}^5 f_i x_i}{\sum_{i=1}^5 f_i} = \frac{1300}{50} = 26$$

Using short-cut method

Marks	f	x	$d = \frac{x-25}{10}$	f d
0-10	6	5	-2	-12
10-20	9	15	-1	-9
20-30	17	25=A	0	0
30-40	10	35	+1	+10
40-50	8	45	+2	+16
	$\sum f = 50$			$\sum fd = 5$

Now
$$\bar{X} = A + \frac{\sum fd}{N} \times h$$

$$\begin{aligned}\bar{X} &= 25 + \frac{5}{50} \times 10 \\ &= 26\end{aligned}$$

2 First we shall calculate the cumulative frequency distribution

Class	f	Cumulative Frequency
0-20	5	5
20-40	15	20
40-60	30	50
60-80	f_4	$50+f_4$
80-100	8	$58+f_4$

Median class is 40-60 since median value is given 50,

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

$$50 = 40 + \frac{\left(\frac{58+f_4}{2} - 20\right)}{30} \times 20$$

$$50 - 40 = \frac{58 + f_4 - 40}{3}$$

$$18 + f_4 = 30$$

$$f_4 = 30 - 18 = 12$$

3 We have

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

$$35.4 = 3 \text{ Median} - 2 (38.6)$$

$$3 \text{ Median} = 35.4 + 77.2$$

$$\text{Median} = 112.6/3 = 37.53$$

Check Your Progress 4

1 After arranging the given data in ascending order of magnitude, we have

10, 10, 11, 11, 12, 12, 12, 13, 13, 14, 14, 14, 14, 15, 15, 15, 15, 15, 16,
16, 16, 16, 17, 17, 17, 17, 18, 18, 19, 19, 19, 19, 20, 20, 21, 21, 22, 23,
25, 25

If you want, you can also present it in **descending order, table form**, or calculate **mean / median / frequency**.

The lowest or smallest observation = $x_s = 10$

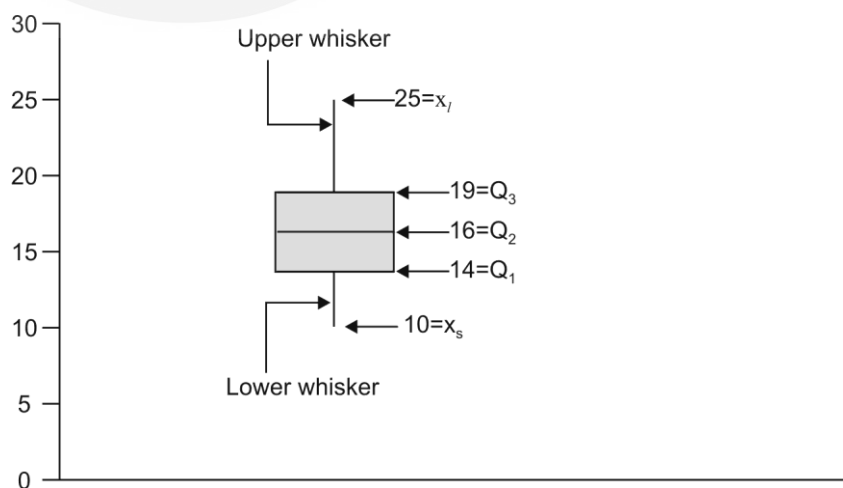
$$\begin{aligned}\text{First quartile } (Q_1) &= \left(\frac{40+1}{4} \right)^{\text{th}} \text{ item} \\ &= 10.25^{\text{th}} \text{ item} = 10^{\text{th}} \text{ item} + 0.25 (11^{\text{th}} \text{ item} - 10^{\text{th}} \text{ item}) \\ &= 14 + 0.25 (14 - 14) = 14\end{aligned}$$

$$\begin{aligned}\text{Second quartile } (Q_2) &= 2 \left(\frac{40+1}{4} \right)^{\text{th}} \text{ item} \\ &= 20.5^{\text{th}} \text{ item} = \text{Mean of } 20^{\text{th}} \text{ and } 21^{\text{st}} \text{ items} \\ &= \frac{16+16}{2} = 16\end{aligned}$$

$$\begin{aligned}\text{Third quartile } (Q_3) &= 3 \left(\frac{40+1}{4} \right)^{\text{th}} \text{ item} \\ &= 30.75^{\text{th}} \text{ item} = 30^{\text{th}} \text{ item} + 0.75 (31^{\text{st}} \text{ item} - 30^{\text{th}} \text{ item}) \\ &= 19 + 0.75 (19 - 19) = 19\end{aligned}$$

The largest observation = $x_l = 25$

Using above calculations box plot based on five-number summary (i.e. smallest observation x_s , first quartile (Q_1), second quartile (Q_2), third quartile Q_3), largest observation x_l) is given below:



2 After arranging the given data in ascending order of magnitude, we have

22, 25, 26, 26, 27, 27, 28, 31, 31, 33, 33, 34, 37, 39, 42, 44, 46, 48, 50

The lowest or smallest observation = $x_s = 22$

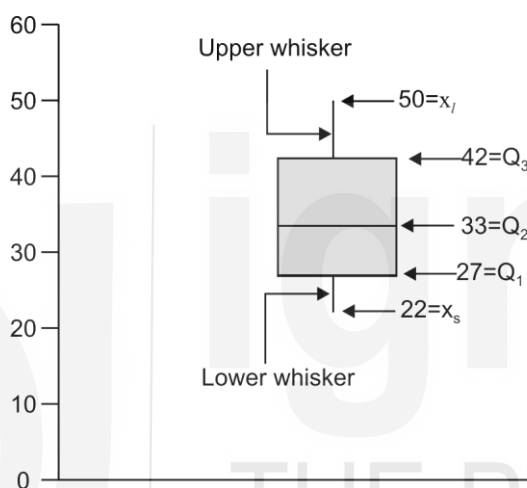
First quartile (Q_1) = $\left(\frac{19+1}{4}\right)^{\text{th}}$ item = 5th item = 27

Second quartile (Q_2) = $2\left(\frac{19+1}{4}\right)^{\text{th}}$ item = 10th item = 33

Third quartile (Q_3) = $3\left(\frac{19+1}{4}\right)^{\text{th}}$ item = 15th item = 42

The largest observation = $x_l = 50$

Using above calculations box plot based on five-number summary (i.e. smallest observation x_s , first quartile (Q_1), second quartile (Q_2), third quartile Q_3), largest observation x_l) is given below:



2.9 REFERENCES

IGNOU Course Material

1. Post-Graduate Diploma In Applied Statistics (PGDAST), MST 001: Matrices, Determinants and Collection of Data, UNIT 13: Classification of Data (for Sections 2.2, 2.3 and related portions in Sections 2.0, 2.1, 2.7, 2.8 and Check Your Progress questions)
2. PGDAST, MST 001: Matrices, Determinants and Collection of Data, UNIT 15: Graphical Representation of Data-I (for Section 2.4, 2.4.1 and related portions in Sections 2.0, 2.1, 2.7, 2.8 and Check Your Progress questions)
3. PGDAST, MST 001: Matrices, Determinants and Collection of Data, UNIT 16: Graphical Representation of Data -II (for Sections 2.4.2, 2.6.4 and related portions in Sections 2.0, 2.1, 2.7, 2.8 and Check Your Progress questions)
4. PGDAST, MST 002: Descriptive Statistics, UNIT 1: Measures of Central Tendency (for Sections 2.5, 2.6 to 2.6.3 and related portions in Sections 2.0, 2.1, 2.7, 2.8 and Check Your Progress questions)

Reference Book

5. "Introduction to Probability and Statistics for Engineers and Scientists", 3rd Edition, Sheldon M. Ross, Elsevier Academic Press, 2004