UNIT 1 DESCRIPTIVE STATISTICS

Structure **Page No.**

1.1 INTRODUCTION

Most of us associate 'statistics' with the bits of data that appear in news reports: Cricket batting averages, imported car sales, average high temperature on a particular day etc. Advertisements often claim that data show the superiority of the advertiser's product. The word statistics, which is derived from the word 'state', entered the English vocabulary in the eighteenth century. It was used then, and still is used, to mean one or more sets of numerical data on various items like population, taxes, wealth, exports, imports, crop production, etc., which are of interest to state officials. There are two ways to use the word statistics. If we say 'statistics is', we are generally referring to the science of statistics. If we say 'the statistics are', we are referring to numbers such as batting averages, the number of unemployed during the month of October, or the number of deaths from malaria during a given year. It is hard to come up with a concise definition of statistics because it is a broad subject that has many facets. Commonly, it is believed that statistics involves the collection, organisation, analysis, and interpretation of data.

There are several reasons why the scope of statistics and the need to study the subject statistics have grown enormously in the last fifty years. One reason is the increasingly quantitative approach employed in all the sciences, as well as in business and many other activities which directly affect our lives. This includes the use of mathematical techniques in the evaluation of anti-pollution controls, in inventory planning, in the analysis of traffic patterns, in the evaluation of teaching techniques and so forth. The other reasons are that the volume of data that is collected, processed and disseminated to the public for one reason or the other has increased almost beyond comprehension. More and more persons with some knowledge of statistics are therefore needed to take an active part in the collection and analysis of the data, as also in all of the preliminary planning. One question that naturally arise at this stage is why and how such large volume of data should be collected, organised and analysed? We shall address such questions in this unit.

We shall begin by discussing the need for collecting data. We shall then talk about different types of data and the ways of arranging them to obtain any logical conclusions from them. Graphical methods of presenting data are also discussed. Finally, we shall discuss various measures that are commonly used to summarise information contained in a data set.

Objectives: After studying this unit you should be able to

- organise the given set of data in a meaningful way;
- construct a frequency distribution of a variable from a given set of data;
- represent graphically a frequency distribution of a variable and interpret the information suggested by it;
- obtain the 'average level' of a given set of data where its frequency distribution is centred by calculating its mean, median or mode;
- obtain the degree to which the individual measurements in a given data vary about this average by calculating its variance and standard deviation.

1.2 COLLECTING DATA

We start with a few examples which provide you a general idea about situations where we need to handle a large amount of data and where statistics can play a significant role. In these examples, we try to raise issues, which can be handled adequately by the various statistical tools with which we will be introduced to as we go along in this course.

Example 1: Suppose we are in the process of drawing up a comprehensive plan for developing public medical facilities in a big city. To work out the rates to be charged for various services proposed to be offered by the hospitals to be set up under the plan, it is necessary to know the economic conditions of the one million households constituting the entire **population** of the city. For this purpose, we need to divide the households which are the **individuals** about which information is sought, into three broad categories - High Income Group (HIG), Middle Income Group (MIG) and Low Income Group (LIG) - according to certain criteria. How do we proceed? Obviously, it would appear as if the task would involve visiting each of the one million households to enquire about, say their monthly incomes. What are the issues and difficulties in implementing this proposal?

Population is a collection of all the individuals we are studying

There are many; for instance:

a) visiting one million households - and in many cases, repeat visits may have to be made following non-availability of respondents at the time of visit, is a time consuming affair and we may not have enough time at hand;

b) to cover such a large number of households, we may have to employ a very large number of investigators and that will mean a lot of expenditure;

c) even if we are in a position to afford the time and money that we need to cover all the households, two more issues crop up:

   i) As the survey may take a long time the figures on monthly income of households covered in the beginning of survey and those of the households covered later may refer to different time periods.

   ii) the data that we will be collecting will be a series of one million figures and obviously such a series as it is will hardly make any sense. We would need proper methodology to properly compile, to analyse and to interpret the data so as to make some sense out of these.

Sample is a collection of some, but not all, of the individuals of the population under study, used to describe the population

In view of (a) and (b), it seems that the process of complete enumeration of all households will not be really efficient and we will have to think of an alternative procedure. To obviate difficulties mentioned in (a) and (b), we might consider only a subset of the population viz., a **sample** of households and collect information on their earnings only. Then try to use the fractional information thus collected to make a guess about the actual state of affairs pertaining to the totality of households.

Now, of course, this method will reduce the cost and time required for the survey and we will have a shorter series of figures. However, issues enumerated in (c) remain to be handled. Moreover, several other new issues like the following would come up now,

d) how do we select our sample?

e) how do we relate the sample findings to the state of affairs in respect of the entire population of one million households?

f) also, even if we are in a position to make such an assessment on the basis of the information contained in the sample, how much reliability should we attach to it, since after all the assessment is based on our observations in a single sample which is perhaps comprised of a small portion of the totality and the same could be quite different if we had a different sample. This aspect surely introduces an element of 'uncertainty' in our conclusion as different samples may differ in respect of their information contents and thus may lead to different conclusions relating to the same population. How do we handle the situation then?

* * *

Thus, though the problem initially looked rather simple, the issues involved are not so by any standard. Solutions of such issues come under the purview of statistics.

The population i.e. the totality of households in the above example is finite and observable. However, often the population is neither finite nor physically existent; rather, it may only be conceivable or hypothetical.

Example 2: Consider a system where 'customers' arrive at a 'counter' for 'service' 'Customers' may be patients coming to a clinic for medical attention or may be aircrafts waiting for clearance from the air traffic control to take off or even broken down machines in a factory waiting for the attention of an operator and so on. Our objective is to prescribe a policy so that congestions can be avoided. However, neither the number of arrivals is fixed on all occasions nor is the service time the same for all customers - these are usually uncertain and thus subject to chance factors. How do we then propose to proceed?

* * *

Example 3: Suppose a new brand of pain reliever has been marketed recently. The manufacturer claims that it relieves pain 25% time faster than any of the comparable brands already available in the market. How do we propose to verify this claim? Obviously, we have to administer these drugs to a sample of individuals. But then how should the sample be chosen? Also, individuals may react differently to the same drug. How do we take this into account? How do we process the sample data? Finally, the good old question - to what extent can we generalise our sample findings so as to be able to come up with a conclusion pertaining to the entirety? How reliable is the sample finding in this case?

* * *

Example 4: The yield of a certain crop is dependent on the location of the plot, amount of fertilisers applied, amount of rainfall, availability of irrigation facilities etc. However, it is also known that even when all these factors are applied equally, the yields may still vary. This is so because the factors we listed may not exhaust all the factors that influences the yield. Given such a situation can we build up an appropriate forecasting formula involving the identifiable factors so that the yield can be predicted adequately? Even if we can come up with such a formula, can we judge how good the formula is?

* * *

Many such examples may be cited to illustrate the areas of application of statistics. However, it will be gradually apparent that the basic issues involved in these illustrations are similar and can be discussed within a broad framework and this framework is provided by statistics.

To study a given phenomenon, the basic raw material would be data (information) relating to it. For example to study the growth in the use of telephones, we may collect

the number of telephones that several workers install on a given day or that one worker installs per day over a period of several days. These figures then constitute our data for studying growth in the use of telephone.

Observation before it is arranged and analysed is called raw data. For data to be useful, our observations need to be organised so that we can pick out trends and come to logical conclusions. We shall thus discuss the techniques of arranging data in tabular and graphical forms to be able to make genuine sense out of it. First, we shall start with describing different kinds of data.

E1) Cite at least two examples from your own experience illustrating the application of statistics

---

1.2.1 Kinds of Data

The operation of collection of relevant data comes in the initial phase of any statistical study. Data relevant for a study can be obtained either from published works or from the collections of the government or research organisations or through direct fieldwork. The mode of collection of data and the methodology for analysing the same comprise the core of statistical discipline. Statisticians gather data from a sample. They use this information to make inferences about the population that the sample represents. Thus sample and population are relative forms. A population is a whole, and a sample is a fraction or segment of that whole.

Whenever the data are numerical, then the corresponding characters are called **variables**. Thus, the number of items failing to meet specifications in a lot of 100 items, the daily number of customers visiting a particular shop, the hourly number of telephone calls received by an operator, the life (in hours) of an electric bulb etc. are all examples of variables.

In some cases, however, the data may not be numerical in nature. This will be the case when, for example, one is examining a lot of manufactured items and classifying them as either 'good' or 'bad'. Similarly, if each member of a group of individuals is asked to say 'yes' or 'no' according as his/her monthly income is at least Rs.5000/- or less than Rs.500/- a month, the resulting data will not be numerical. This exercise will produce only a series of "yes"es and "no"s. However, in both these instances, the data are easily convertible to numerical terms. For instance, in the last example, we may code a "yes" as 1 and and a "no" as 0. Such characteristics (e.g. quality of manufactured items or the salary position) are called **attributes**.

You may note here that in the first three examples given above, the variables take on only some isolated values; for instance, the number of items failing to meet specifications in a lot of 100 items can be 0 or 1... or at most 100, i.e., a whole number between 0 and 100, and never a figure like 3.7, say. Such variables which take on only isolated values (which are often integers, but need not be such always) are called **discrete** or **discontinuous** variables. This kind of data result from counts and therefore, values jump from one point to the next with no possible measures in between. On the other hand, certain variables are such that they may take on any value along a suitable scale. For instance, the distance (in meters) between two points, in the interval between 210.5 and 320.6 can take any value like 210.56, 210.687, 315.685 and so on. Such variables are called **continuous variables**. Variables representing height, weight, age, time to complete a journey, temperature etc. are all of this variety. Although continuous data can take on a theoretically infinite number of possible measures, the values that we use in practice are determined by

i) the precision of our measuring instrument and;

ii) how precise we need to be.

For example, if you measure the length of a new born baby, you would record it as 48

cm., 49 cm., etc. For most practical purposes we will not be required to measure a baby's length to the fifth decimal place.

We now present some data to illustrate the concepts of attributes and variables:

Example 5: The following table is a summarised version of data relating to the educational background at graduation level of students admitted in the MBA Programme of a certain college. Here the educational background of students admitted is an attribute which can be either arts, science, commerce, engineering or any other branch.

Table 1: Educational Background of Students Admitted to the MBA Programme in a College

Graduation Background	Year				
	1994-95	1995-96	1996-97	1997-98	1998-99
Arts	5 (4.0)	7 (7.1)	10 (10.7)	15 (10.1)	18 (11.25)
Science	10 (8.0)	5 (5.1)	12 (8.1)	17 (11.5)	10 (6.25)
Commerce	5 (4.0)	3 (3.1)	12 (8.1)	16 (10.8)	12 (7.5)
Engineering	103 (84.0)	83 (84.7)	109 (73.1)	98 (66.2)	118 (73.75)
Others	-	-	-	2 (1.4)	2 (1.25)
Total enrolled	123	98	149	148	160

(Figures in parentheses are percentages of the total number of students enrolled)

* * *

Example 6: The following table gives the raw data relating to the marks (out of 10) of 100 students in a statistics examination.

Table 2: Marks of 100 Students in a Statistics Examination

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

Here the marks in the test is a variable, which varies from one student to another. Since the marks are in whole numbers between 0 and 10, it is a discrete variable. The data in the given format is called **ungrouped data**.

* * *

The way the marks have been reported does not help us in knowing who scored how much. Perhaps that is not important for the purpose of the study at this stage. We are interested in knowing how the performance was in general? Was the test very simple in the sense that a large proportion of students scored high marks? Was the test too difficult in the sense that a large proportion of students scored very low marks? It may not be easy for you at this moment to answer such questions with the data available.

To respond to such questions, we have to present the data in an organised manner. We shall take up this example again in the next section and try to find answers to the above questions but before that, let us look at the data corresponding to continuous variables.

Example 7: Consider the following data, which relate to life (in hours) of 100 electric bulbs.

Table 3: Lives of 100 Electric Bulbs

511.6	977.7	600.2	1099.7	803.7
923.4	1108.3	906.7	759.6	1111.9
918.3	1051.1	992.5	817.2	665.3
1143.6	948.4	939.8	1163.0	715.2
936.1	750.5	991.2	1199.5	950.2
1061.7	1027.7	995.1	966.5	1146.5
848.0	956.8	1100.0	955.2	1023.0
900.5	982.3	699.2	1069.8	1245.3
1059.5	1091.0	850.7	1219.3	1012.6
1053.2	939.5	777.8	749.6	980.8
991.3	1016.3	930.4	1242.2	1131.4
1314.7	1137.2	763.1	1394.4	117.3
1204.1	980.1	922.3	1057.7	907.2
808.0	857.7	1127.1	934.3	1262.3
965.4	873.4	955.1	806.5	1033.0
1068.3	950.3	930.6	1000.1	898.5
1293.1	940.9	1293.8	1035.2	706.0
880.9	912.2	803.5	922.6	846.1
1092.3	1182.0	985.2	945.3	835.0
1001.5	1048.8	895.1	1067.2	1062.8

Here the life of a bulb is a variable which varies from one bulb to another. In this example, the figures have been recorded correct to 1/10th of an hour. All values larger than 511.55 but not larger than 511.65 have been approximated and represented by 511.6. Although, conceivably the life could be any value on the continuous scale from zero to infinity, the limitation on the part of the measuring instrument invariably imposes an artificial discreteness in the data. This is due to limitations of measuring instruments. For instance, in this case the scale could measure only up to the first place of decimal; but, if we had a scale which could measure up to, say five places of decimals, the accuracy would have increased. No matter how fine the scale is, one would have to stop after a finite number of places after the decimal point and as such a discreteness would any way creep in. But, theoretically, the variable under consideration is continuous in character as it can adopt any value over a specified interval, finite or otherwise.

* * *

You may now try the following exercises

E2) Identify each of the following as a population or as a sample.

a) All the adult males residing in India.

b) Twenty cancer patients chosen to participate in a program to test a new drug.

c) All the AIDS patients who could conceivably be given a new treatment for the disease.

E3) Data on the variables below were recorded for a study in a school in Delhi. Which of these are continuous, and which are discrete?

a) age

b) year of birth

c) height

d) number of students admitted to the school in a calendar year

e) weight

E4) Give two examples of situations that would yield continuous data.

Let us now see how the raw data in Examples 6 and 7 can be arranged to obtain any logical conclusion from them and obtain an answer to the questions asked earlier.

1.2.2 Frequency Distribution of a Variable

From the data in Example 6 for instance, it is not possible to readily answer questions like how many students scored 3 and above or how many of them scored between 5 and 7 or what is the score obtained by most of the students. It will be easier to answer such questions if we compress the data in the form of a Table 4 given below. Here we put a tally mark against a score as soon as we come across the same as we go along the list (data) in any given order. For instance, if we read the data column wise then first entry in first column is 2, so we put a tally mark against 2 as shown in Table 4. Number of tally marks against 2 indicate the number of students getting 2 marks. You may observe here in this table that every fifth tally is a slash over the first four tally marks.

Table 4: Tally marks against marks scored

Scores	Tally Marks
0	/
1	//
2	////
3	𝍸 ////
4	𝍸 𝍸 𝍸
5	𝍸 𝍸 𝍸 𝍸 /
6	𝍸 𝍸 𝍸 𝍸
7	𝍸 𝍸 𝍸
8	𝍸 ////
9	////

Finally we count the tallies for each variable value (i.e. score) to obtain the number of times it appears in the data set. We call this number the **frequency** of the variable.

Definition: The frequency of any value occurring in a data set is the number of times the value occurs in the set.

Once we know the frequency with which the values occur in a data set, we can construct a table showing the frequency of each value against it. We call this table a **frequency distribution**. Table-5 gives the frequency distribution for the data set in Example 6.

Table-5: Frequency Distribution of marks of 100 students

Scores	Frequency	Relative Frequency
0	1	.01
1	2	.02
2	4	.04
3	9	.09
4	15	.15
5	21	.21
6	20	.20
7	15	.15
8	9	.09
9	4	.04
Total	100	1.00

Table 5 above gives a **grouped version** of the data in Table 2. An entry in the second column gives the number of students receiving the corresponding score that is depicted in the same row under the first column. Thus, there are 10 groups (**classes** or

categories) in the table above. Each group corresponds to a single value of the underlying variable. Note that the third column gives the proportion of students receiving the various scores called the **relative frequency** of the value.

Definition: Relative frequency of a value occurring in a data set is the frequency of a value as a fraction or a percentage of the total number of observations.

Now suppose we want to see how many students got less than 6 marks. Then we have to add up the frequencies of 0,1,2,3,4 and 5 marks to obtain the number of students getting marks less than 6. This gives us the **cumulative frequency** of the **'less than'** type of 6. The cumulative totals of the frequencies obtained by proceeding from the top class of the table downwards are called the 'less than' type cumulative frequencies. Similarly, if we add up the frequencies from the bottom class upwards, we get the cumulative frequencies of the '**more than**' type. We have given both these types in Table 6 for the data under consideration.

Table 6: Cumulative Frequency of the marks of 100 students:

Score	Cumulative Frequency (Less than type)	Cumulative Frequency (More than type)
0	1	100
1	3	99
2	7	97
3	16	93
4	31	84
5	52	69
6	72	48
7	87	28
8	96	13
9	100	4

For instance, the less than type cumulative frequency of 5 is 53. This means that 53 students scored 5 or less marks. Also, the greater than type cumulative frequency of 5 is 68. Thus, there are 68 students who scored 5 or more marks. From the table, we can also see that the number of students scoring above 5 but not above 8 is 96-53 = 43. We call the data presented in Table 6, grouped. This is because we have grouped together all the observations having the same value.

We would like to mention here that the operation of grouping the observations is not unique. For example, consider the following grouping of the set of observations with the values in a group as given in Table 7 below.

Table 7: Frequency Distribution of marks of 100 students

Scores	Frequency	Cumulative Frequency	
		Less than type	More than type
0-1	3	3	100
2-3	13	16	97
4-5	36	52	84
6-7	35	87	48
8-9	13	100	13
Total	100		

The grouping in Table 7 informs us that 35 students scored 6 or 7 marks in the test, but it does not give us the exact number of students receiving each of these marks. This shows that some information has been lost in this process of classifying observations into classes. Of course, as we have seen, this problem is avoidable in case of discrete variables by making each class correspond to a single value of the variable provided, the number of distinct values that the variable assumes is not too large as is the case with Example 6 discussed above.

But, summarisation of data at the expense of loss of such information becomes almost unavoidable in the case of a continuous variable. In this case if we take a class corresponding to each different value taken by the variable, then the resulting number of classes will be unduly large. This approach will then look artificial since a continuous variable, as you know, by definition is capable of assuming any of the values represented in a relevant interval. In such a case, we decide on the mode of classification depending on the nature of the data and the purpose of the study. Additionally, the following points are of help and can be taken care while deciding on the classes;

i) each class should correspond to an interval of values of the variable;

ii) the classes should be non-overlapping and exhaustive i.e. an observation must be included in exactly one of the classes;

iii) the number of classes should not be too small since otherwise the actual nature of the distribution may be difficult to visualise and thus the summarisation will fail to bring out the actual characteristics of the distribution;

iv) the number of classes should also not be too large;

v) classes should preferably be of equal width, since otherwise the frequencies of various classes may not be comparable.

As a rule of thumb, the number of classes should be between 10 and 20 wherever the total frequency is more than 1000. Let us now take up the data in Table 3, corresponding to Example 7 to illustrate the points made above. You may observe here that the smallest and the largest values are 511.6 and 1394.4 respectively. Here we take classes as given in Table 8 below.

Table 8: Tally marks of lives of 100 electric bulbs

Life (hours) (inclusive of end point)	Tally marks
510.6-590.5	/
590.6-670.5	//
670.6-750.5	𝍸
750.6-830.5	𝍸 ///
830.6-910.5	𝍸 𝍸 ///
910.6-990.5	𝍸 𝍸 𝍸 𝍸 𝍸 /
990.6-1070.5	𝍸 𝍸 𝍸 𝍸
1070.6-1150.5	𝍸 𝍸 //
1150.6-1230.5	𝍸 /
1230.6-1310.5	𝍸 /
1310.6-1390.5	/

Note that here the values have been recorded correct to 1/10 th of an hour. The life x hours of a bulb with $510.55 < x \leq 510.65$ has been recorded as 510.6; the class 510.6 - 590.5 actually includes all bulbs with life x hours satisfying $510.55 < x \leq 590.55$. While 510.6 and 590.5 are respectively called the lower and upper **class limits** 510.55 and 590.55 are referred to as the lower and upper **class boundaries**. Similar is the case with the other classes. Once the classes have been determined, the frequency distribution can now be obtained exactly as before.

The difference between the upper and lower boundaries of a class is called the **class interval**. The mid point of a class is called its **class mark**. Thus $(510.55 + 590.55)/2 = 550.55$ is the class mark of the first class. The last column of Table 9 gives **frequency density** of a class. It is the frequency per unit length of the class interval. For the first class, frequency density is $\frac{1}{590.55 - 510.55} = 0.0125$ and so on.

Table 9: Frequency Distribution of Lives of 100 Electric Bulbs

Life (hours): Class boundaries (exclusive of left boundaries but inclusive of right boundaries)	Frequency	Relative frequency	Cumulative frequency		Frequency density
			'Less than' type	'More than' type	
510.55-590.55	1	.01	1	100	0.0125
590.55-670.55	2	.02	3	99	0.025
670.55-750.55	5	.05	8	97	0.0625
750.55-830.55	8	.08	16	92	0.1
830.55-910.55	13	.13	29	84	0.1625
910.55-990.55	26	.26	55	71	0.325
990.55-1070.55	20	.20	75	45	0.25
1070.55-1150.55	12	.12	87	25	0.15
1150.55-1230.55	6	.06	93	13	0.075
1230.55-1310.55	6	.06	99	7	0.075
1310.55-1390.55	1	.01	100	1	0.0125
Total	100	1.00			

And now some exercises for you.

E5) The number of nurses on duty each day at a hospital are grouped into a distribution having the classes 20-34, 35-49, 50-64, 65-79 and 80-94. Find

a) the class limits

b) the class boundaries

c) the class marks

d) the class interval of the distribution.

E6) The total scores X obtained by 50 students in a psychology test of 100 marks are given below.

75	89	66	52	90	68	83	94	77	60
38	47	87	65	97	49	65	70	73	81
85	77	83	56	63	79	69	82	84	70
62	75	29	88	74	37	81	76	74	63
69	73	91	87	76	58	63	60	71	82

Answer the following question on the basis of the data given above.

a) Is the random variable X = Score of a student, discrete or continuous? What are the minimum and maximum scores?

b) Using the classes $20-29, 30-39, 40-49, \ldots$ and $90-99$ draw up the frequency distribution of X.

c) What percentage of the students score above the pass marks of 50?

d) How many of the students score between 50 and 79?

Consider the following frequency distribution of income of 1000 individuals belonging to a particular section of the population:

Income (Rs.)	Frequency
≤ 1000	40
1000-2000	55
2000-4000	141
4000-6000	152
6000-10000	275
10000-15000	199
15000-25000	103
≥ 25000	35

i) What percentage of people earn more than Rs.4,500?

ii) What percentage of people earn at least Rs.1,500?

iii) What percentage of people have earnings between Rs.2,000 & Rs.5000?

We can thus say that frequency distributions condense large sets of data and display them in an 'easy to understand' form. Graphical methods are also used for presenting data and ideas. Graphical methods provide an effective way to present data as they give an idea of the pattern of variation of the random variable at a glance. We shall now discuss some of the more common forms of graphical presentation and see how they help us to illustrate, clarify and interpret the information contained in the data.

1.2.3 Graphical Representation of Frequency Distributions

A frequency distribution relating to a discrete variable is commonly represented through either of the following two diagrams:

1) **Bar diagrams**: Fig.1 gives the bar diagram for the data given in Table-7.

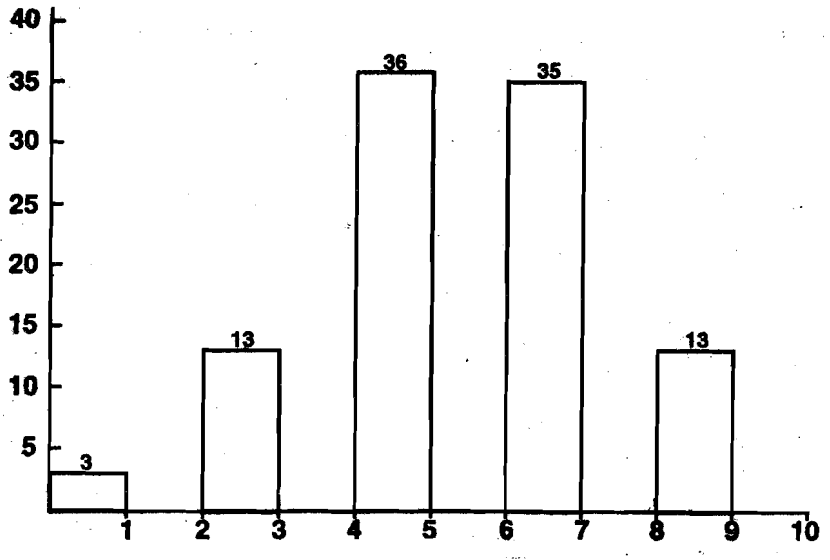

Fig.1

We can easily construct the bar diagram. Here the marks scored by the students are located on the horizontal, x-axis, and the frequencies of their occurrence on the vertical, y-axis. Note that the height of the rectangles, or bars, represent the class frequencies. From the heights of these rectangles you can easily deduce at a glance that a maximum number of students scored between 4-5 marks. For convenience, we have at the top of each rectangle given the corresponding frequency.

2) **Frequency Polygon**: Another form of graphical presentation is the frequency polygon. Frequency polygon for the data in Table 7 for which the bar diagram is drawn in Fig.1 is shown in Fig.2.

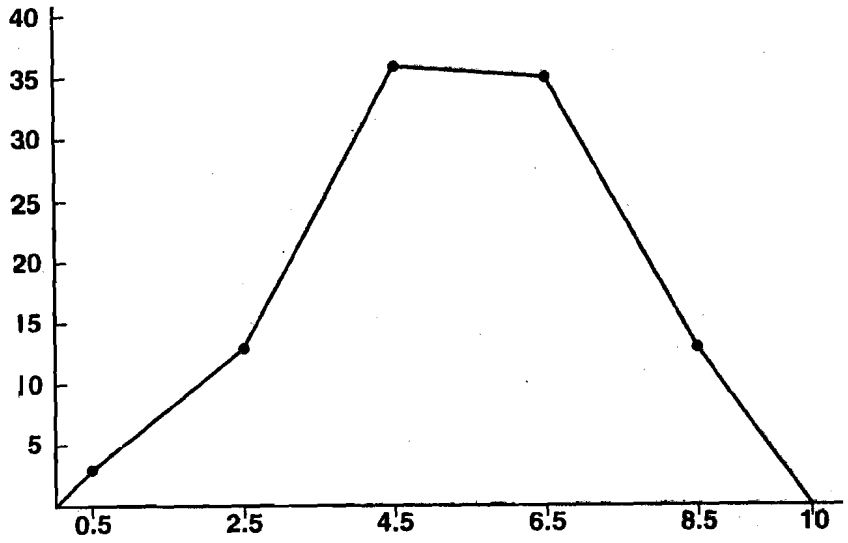

Fig.2

Here the class frequencies are plotted at the class marks and the successive points are connected by straight lines. It is thus tacitly assumed that the frequency of a class interval is concentrated at its mid-point. **Note** that we added classes with zero frequencies at both ends of the distribution to "tie down" the graph to the horizontal scale.

In the case of a continuous variable, one often uses another diagram called a **histogram**. To draw the histogram of a set of given data, take a graph paper with rectangular coordinate axes and plot the class boundaries along the x-axis. Now over each class interval erect a rectangle the height of which equals the relative frequency of this class. The resulting figure is the histogram of the given data. Here we use class boundaries to demarcate the class intervals and not class limits. This ensures that there is no gap left between the rectangles. Thus, the area of these rectangles represent the frequencies of the corresponding classes.

Fig.3 gives a frequency polygon and a histogram for the frequency distribution of lives of 100 electric bulbs given by Table-9.

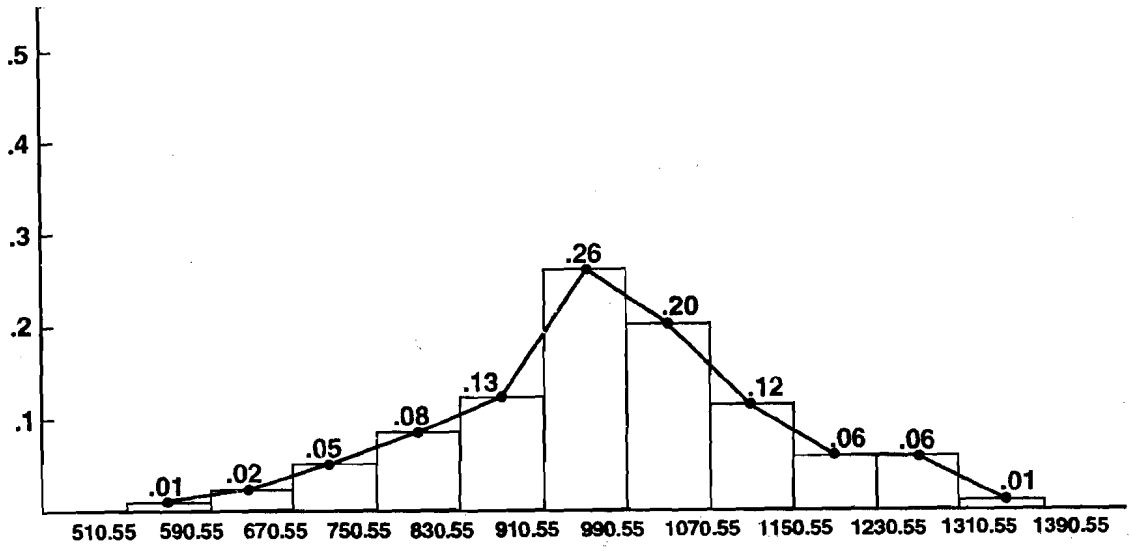

Fig.3

Note that in Fig.3 above we have plotted the rectangles by taking relative frequency of a class along the y-axis. This histogram has the same shape as an absolute frequency

histogram made from the same data set. This is because in both the relative size of each rectangle is the frequency of that class compared to the total number of observations. Presenting the data in terms of the relative rather than the absolute frequency of observations in each case is useful because, while the absolute numbers may change, the relationship among the classes may remain stable. Also note that if the width of the class intervals are the same, then the heights of the rectangles are proportional to their areas.

Histograms and frequency polygons are similar. However, each one has some advantages. In the case of histogram

i) The rectangle clearly shows each separate class in the distribution

ii) The area of each rectangle, relative to all the other rectangles, shows the proportion of the total number of observations that occur in that class. For instance, looking at Fig.3 one can easily conclude that normally, the life span of an electric bulb is between 910-990 hrs.

Similarly, frequency polygon has certain advantages viz.,

i) The frequency polygon is simpler than its histogram counterpart

ii) It sketches an outline of the data pattern more clearly

iii) The polygon becomes increasingly smooth and curve like as we increase the number of classes and the number of observations.

Representation of a frequency distribution graphically on the basis of cumulative frequencies is also quite common. The diagram that one gets by plotting cumulative frequencies (less than type) against the upperclass boundaries and joining the points by line segments is classed the **less than type ogive** of the frequency distribution. Similarly one gets the **more than type ogive** by plotting the more than type cumulative frequencies against lower class boundaries. Less than type ogive for the frequency distribution of lives of 100 electric bulb given by Table-9 is given in Fig.4. Such figures are useful for finding how many measurements are located above or below a given point.

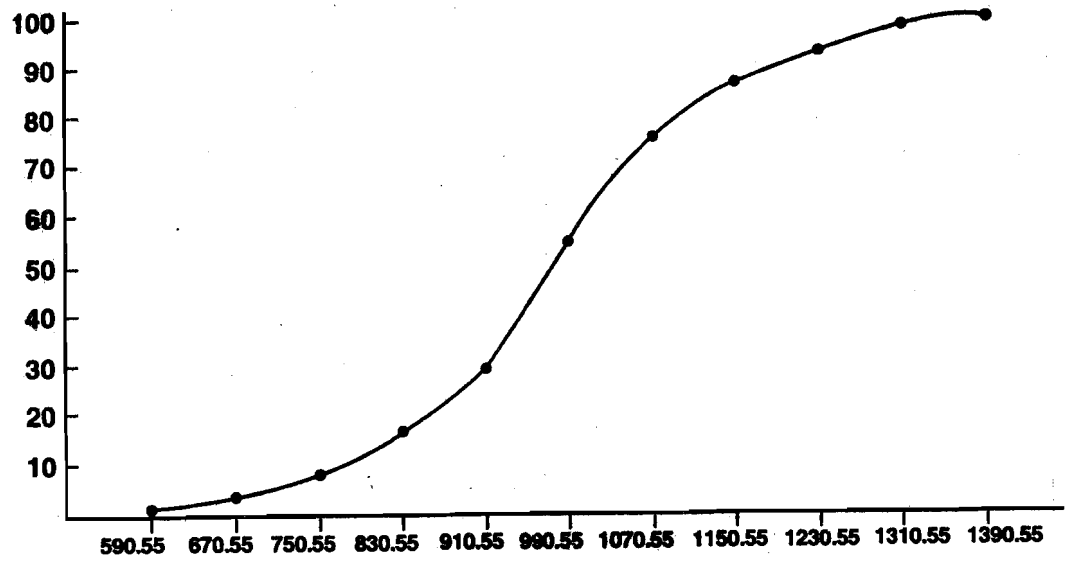

Fig.4

Frequency distributions are often presented graphically as **pie charts**, where a circle is divided into sectors, pie-shaped pieces, which are proportional in size to the corresponding frequencies or percentages. To construct a pie chart, we first convert the class frequencies into percentages of the total number of observations. Then since a complete circle corresponds to 360 degrees, we obtain the central angles of the various sectors by multiplying the percentages by 3.6. Fig.5 gives the pie chart for the data in

Table 7 giving the marks scored by 100 students in a statistics test. In this table you may observe that the total frequency is 100. Hence, the frequencies given by column 2 of this table also equals the percentage frequencies. Regions shaded differently in Fig.5 represent these percentages.

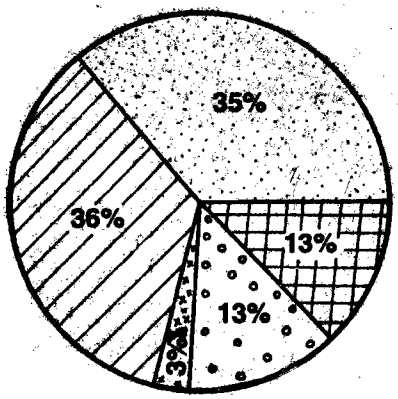

Fig.5

You may now try the following exercises:

E8) In a sample of 60 families in a certain locality, the number of children per family are recorded as follows:

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

Obtain a frequency distribution of the these observations and represent it in a suitable diagram. Also draw the cumulative frequency graph of the above data. What proportion of families will have at least 2 children? Also compute the percentage of families having not less than 2 and not more than 4 children.

E9) The yields (in quintals) of a grain from 500 small plots grouped in classes with a common width of the class interval are given below.

Class boundaries	Frequency
2.7-2.9	4
2.9-3.1	15
3.1-3.3	20
3.3-3.5	47
3.5-3.7	63
3.7-3.9	78
3.9-4.1	88
4.1-4.3	69
4.3-4.5	59
4.5-4.7	35
4.7-4.9	10
4.9-5.1	8
5.1-5.3	4

Represent the data graphically. Also draw the cumulative frequency graph.

E10) Data given below show the areas of the various continents of the world in million square miles.

Continent	Area (million square miles)
Africa	11.7
Asia	10.4
Europe	1.9
North America	9.4
Oceania	3.3
South America	6.9
Russia and other former Soviet Republics	7.9

Represent the above data by means of a suitable diagram.

So far, we learned to construct tables and graphs using raw data. We see that the frequency distribution of a variable summarises the statistical data on the variable and brings out the pattern and feature of its variation. But what if we need more exact measures of a data set? In that case, we can use single numbers, called summary statistics to describe certain characteristics of a data set. For instance, how do you compare the performance of two batsmen in the game of cricket? You must have seen people saying that Batsman-1 is better because his average score is better; hardly, you'll hear anybody making a match to match comparison. Thus the average number of runs scored is a summary of all his scores in all the matches he has played so far. So in this case average is the summary statistics that can be used to rate the performance of the batsmen.

We shall now discuss the summary measures that are commonly used to summarise information contained in a data set.

1.3 SUMMARISATION OF DATA

The choice of a single number or summary statistics that we choose to summarise a given data depends on the particular characteristics we want to describe. In one study we may be interested in the value which is exceeded by only 25 percent of the data; in another, in the value which exceeds the lowest 10 percent of the data; and in still another, in a value which describes the centre or middle of the data. The statistical measures which describe such characteristics are called **measure of location or measures of tendency**. Some of such important measures are:

i) measures of central tendency

ii) measures of dispersion

We shall now discuss these measures one-by-one.

1.3.1 Measures of Central Tendency

If you take a close look at any data set, you would notice that though the manifestation of the variable is different for different observational units, the values tend to cluster around a central value. This property is referred to as **central tendency**. For instance, look at the data sets in Table 2 and Table 3. In these cases, at a first glance, the corresponding observations seem to be clustering around 6 and 1000 or thereabouts respectively.

A representative value around which a given set of observations tends to cluster (or equivalently be located) is a measure of central tendency or location or is simply an average. **Arithmetic mean** (a.m.), **median** and **mode** are the three commonly used averages. Other averages are **geometric mean** and **harmonic mean**.

Arithmetic Mean

The most popular measure of central location is what the layman calls an 'average' and what the statistician calls an **arithmetic mean**, or simply a **mean**. The arithmetic mean

is the sum of the numbers included in the relevant set of data divided by the number of such numbers.

Mean from Ungrouped Data

Let there be N items or units in a population. In Table-2, N=100. If we order these numbers from 1 to N, x_1 being the first number, x_2 being the second number, and so on up to x_N, which is the Nth number, then the population mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

In particular, if we have N=4, where $x_1 = 30, x_2 = 50, x_3 = 70, x_4 = 90$, then

$$\bar{x} = \frac{30 + 50 + 70 + 90}{4} = 60$$

Thus, arithmetic mean in this case turns out to be 60.

E11) The average weekly wage of 60 workers in a factory was calculated as Rs.80.50. Later, it was found that for one worker the wage was recorded as Rs.92.00 whereas the actual figure should have been Rs.80.00. What is the corrected average weekly wage?

E12) In a construction project, 60% of the labourers work on a daily rate of Rs.60.00 a day. The rest are paid at a weekly rate of Rs.450.00. If the rate for the first group is increased by 15% and the rate of the second group is reduced by 20%, will the average income of the labourers increase or decrease? (Sundays are holidays).

Mean from Grouped Data

The computation of the mean is easy whenever data are grouped in such a way that each class corresponds to single observed value of the variable (see Table 5). For example, suppose that the distinct values in the data are $x_1, x_2, \ldots, x_k$ with $f_1, f_2, \ldots, f_k$ as the corresponding frequencies (and thus k is the number of classes). Since each x_i appears f_i times, $i = 1, 2, \ldots, k$ in the data, the sum of all observations is equal to

$$\underbrace{x_1 + x_1 + \ldots + x_1}_{\longleftarrow\ f_1 \text{times} \longrightarrow} + \underbrace{x_2 + x_2 + \ldots + x_2}_{\longleftarrow\ f_2 \text{times} \longrightarrow} + \underbrace{x_k + x_k + \ldots + x_k}_{\longleftarrow\ f_k \text{times} \longrightarrow}$$
$$= x_1 f_1 + x_2 f_2 + \ldots + x_k f_k$$

Now the arithmetic mean will be given by

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j f_j \qquad (2)$$

Let us illustrate the above formula through Example 6.

Example 8: Consider the data of Example 6 in the following computational format.

Table 10: Frequency Distribution of Scores of 100 Students in Statistics test

Score (x_j)	Frequency (f_j)	$x_j f_j$
0	1	0
1	2	2
2	4	8
3	9	27
4	15	60
5	21	105
6	20	120
7	15	105
8	9	72
9	4	36
Total	100	535

Thus, $\bar{x} = 535/100 = 5.35$ is the mean marks scored by a student in the test.

How do we calculate the mean of the data given in tables 7 or 9? In either of these tables, it is not possible to calculate the sum of all the observations exactly because from these tables it is not clear which value appeared how many times in the respective data sets. For instance, Table 7 merely tells us that 3 of the observations are either 0 or 1, 13 of the observations are either 2 or 3 etc.. Similarly, Table 9 informs us that there is just one observation which is larger than 590.55 but not larger than 670.55 etc., but it does not inform us the exact numerical values of these observations. Thus, computation of the exact mean is impossible whenever the relevant data is available in the grouped form with more than one variate value in each class or category (e.g. each class in Table 7 represents two variate values and each class in table 9 represents uncountably many values). However, one cannot really complain much, because grouping the observations always leads to loss of such information anyway. In any case, in such situations, only an estimate of the mean can be obtained by making use of Formula (2) above, but with x_j interpreted as the class mark of the j-th class. Observe that here one would implicitly assume that all the observations falling in a given class are located at the class mark. This assumption is not as unrealistic as it may seem. Although all observations cannot be expected to be located at the class marks, some will fall above and some below and then the error of estimation of the mean would be expected to average out. Let us compute the mean of the data given in Table 9.

Example 9: Consider the data of Example 7 in the following computational format.

Table 11: Frequency Distribution of Lives of 100 Electric Bulbs

Life (hours): class boundaries (exclusive of left boundaries but inclusive of right boundaries)	Frequency f_j	Class marks x_j	$x_j f_j$
510.55-590.55	1	550.55	550.55
590.55-670.55	2	630.55	1261.10
670.55-750.55	5	710.55	3552.75
750.55-830.55	8	790.55	6324.40
830.55-910.55	13	870.55	11317.15
910.55-990.55	26	950.55	24714.30
990.55-1070.55	20	1030.55	20611.00
1070.55-1150.55	12	1110.55	13326.60
1150.55-1230.55	6	1190.55	7143.30
1230.55-1310.55	6	1270.55	7623.30
1310.55-1390.55	1	1350.55	1350.55
Total	100		97775

Thus the estimated mean $\bar{x} = 97775/100 = 977.75$ hours.

If the mean of the measurements in each class interval is close to the midpoint of the class interval i.e. the class marks, then the error in approximating $\bar{x}$ by Formula 2 is very small.

You will observe in some cases that the measurements in a sample or a population need not be weighted equally, as in Eqn.(1). For example, suppose that you want to calculate the mean profit rate for a group of firms. Since some firms are much bigger than others, a firms profit rate should be weighted according to its size in determining the average level of profit rates. If w_i is the weight attached to the ith measurement in a sample, the **weighted arithmetic mean** is

$$\bar{x}_w = \sum_{i=1}^{n} x_i w_i \Big/ \sum_{i=1}^{n} w_i \qquad (3)$$

For example, suppose that we have a sample of three firms' profit rates as 10%, 12% and 15%. The firm with the 10% profit rate has assets of 2 crores, whereas the other two firms have assets of 1 crore each. If a firm's assets are used to weight its profit rate, the weighted arithmetic mean of the profit rates of these three firms is

$$\bar{x}_w = \frac{2(10) + 1(12) + 1(15)}{2 + 1 + 1} = \frac{47}{4} = 11.75$$

Thus, the weighted mean is 11.75%.

If you compare Formula (2) with Formula (3) you would notice that the arithmetic mean based on grouped data is a type of weighted arithmetic mean. It is a weighted mean of the midpoints of the class intervals, the weight attached to each particular midpoint being the number of measurements falling within that class interval.

E13) Compute the mean for the data given in E6).

The Median

Another measure of central tendency is the **median**. The median of a given data set is defined to be the middlemost observation or the mean of the two middle observations depending on whether the total number of observations n is odd or even, once the observations are sorted in the increasing or decreasing order of magnitude. Thus there are as many values above the median as there are below.

Median from Ungrouped Data

To calculate the median of a set of n observations, the n observations are arranged in the increasing or decreasing order of magnitude and then $(n + 1)/2$-th observation is identified as the median whenever n is odd; when n is even, then the median is computed as the mean of the n/2-th and (n/2+1)-th observations.

For example, suppose the number of road accidents on the first five days (i.e. Monday through Friday) of a week in a city was 2,7,4,1 and 5. We arrange these observations in the increasing order as 1,2,4,5,7 so that the (5+1)/2-th i.e. the 3rd value which is 4 is the median in this case. On the other hand, suppose we also know that 3 accidents took place on Saturday in the same week so that now we will have six observations which when arranged in the increasing order would look like 1,2,3,4,5,7and the median will be the average of the (6/2)-th i.e. the 3rd and the (6/2+1)-th i.e. the 4th values i.e. (3+4)/2 = 3.5.

Now let us discuss the procedure for computing the median in the case of grouped data.

Median for Grouped Data

For calculating the median from the frequency distribution, consider the data set as represented in the frequency distribution of Table 9. What are the exact observations in the class 510.55-590.55 or, in the next class 590.55-670.55? These are not known from the table. In fact, the individual observations have lost their identities because of grouping which has led to loss of such information. But, unless the observations are known, we cannot really arrange them in order of magnitude. Thus exact calculation of the median in such cases will not be possible and we have to find some way of estimating the same. In such a situation, one proceeds as follows:

The first step in calculating the median from a frequency distribution is to find the class interval that contains the median. To do this, we start with the lowest class interval, cumulate the number of measurements in one, two, three and subsequent class intervals and stop with the interval where the cumulated number of measurements first exceeds or equals n/2. This particular class interval contains the median. Let us illustrate what we have said above with the help of the data given in Table 9. In this case $n/2 = \frac{100}{2} = 50$. If you look in the column with cumulative frequency 'less than type'

then 50 lies in the sixth class interval. Since the cumulative frequency for the fifth class interval is 29 and for the sixth class interval is 55, it is clear that sixth class interval is the first where the cumulative number exceeds 50. Thus this is the class interval in which the median is located. Now to find the median M for the grouped data of this sort, we use the following expression

$$M = \left(\frac{n/2 - c}{f_m}\right) l + L_m \qquad (4)$$

where c = the number of measurements in class intervals below the one containing the median;
f_m = the number of measurements in the class interval containing the median;
l = the width of the class interval containing the median
L_m = the lower boundary of the class interval containing the median.

For the data in Table 9, n=100, c=29, f_m = 26, l=80 and L_m = 910.55 and hence using Formula 4, the median is M = 975.17.

E14) For the following data, calculate an estimate of the median

Class	0-24.9	25-49.9	50-74.9	75-99.9	100-124.9	125-149.9
Frequency	6	11	14	16	13	10

Before proceeding further with the discussion of another measure of central tendency, let us compare the uses of the mean and the median.

Uses of the Mean and the Median

Both the mean and the median are important and useful measures of central tendency. In some circumstances the mean is a better measure than the median, and in others the converse is true. The following factors contributes to the determination of whether the mean or the median should be used.

i) **Sensitivity to Extreme Observations**: The median is often preferred over the mean when the latter can be influenced strongly by extreme observations. Consider for instance, the computation of an average income of the families in an apartment building containing 14 families, 3 of which earn Rs.10,000 per month, 5 of which earn Rs.12,000 per month, 5 earn Rs.15,000 per month and 1 of which earns Rs.1 lakh per month. Then the mean income per month in rupees of the 14 families equals

$$\frac{3(10,000) + 5(12,000) + 5(15,000) + 1,00,000}{14} = \frac{2,65,000}{13} = \text{Rs.}18,929 \text{ approx.}$$

However, this figure is not a very good description of the monthly income level of the majority of the families in the building. A better measure might be the median, which in this case is Rs.12,000 per month. The median is much less affected by the one extreme value.

ii) **Open Ended Class Intervals**: It may happen that in a frequency distribution some intervals do not have finite upper or lower limits. For example, in a frequency distribution of the monthly income of families, two class intervals might be "less than Rs.15,000" and "Rs.30,000 and more". Each of these class intervals is open-ended. With such class intervals,there may be no alternative but to use the median, since calculation of the mean requires a knowledge of the sum of the measurements in the open-ended classes.

iii) **Mathematical Convenience**: The mean rather than the median is often the preferred measure of central tendency because it possesses convenient mathematical properties that the median lacks. For instance, the mean of two

combined populations or samples is a weighted mean of the means of the individual populations or samples. On the other hand, given the medians of two populations or samples, there is no way to determine what the median of the two populations combined or two samples combined would be.

iv) **Extent of Sampling Variation**: Sample statistics such as the sample mean or the sample median are often used to estimate the population mean. A major reason for preferring the mean to the median is that the sample mean tends to be more reliable than the sample median in estimating the population mean. In other words, the sample mean is less likely than the sample median to depart considerably from the population mean. You will be able to appreciate this consideration more when you study estimation later in Block 2.

We shall now move forward and discuss the third measure of central tendency, the mode.

The Mode

Mode is defined as the most frequent observed value of the measurements in the relevant set of data. In a set of observations, if all the observations are distinct so that each of these occur with frequency 1, then it will be meaningless to say each of them is a mode; as such, in such a situation we say that the mode does not exist. However, from the definition, it is clear that a given data set may have more than one mode. For instance, if there are two modes then the set of observations is referred to as a **bimodal** data set. In an interval-grouped data set, the mode is estimated by the class mark of the class depicting highest frequency. In Table 10, the score 5 appears with the highest frequency so that 5 is the mode of the scores of the 100 students. On the other hand, the mode of the data represented in Table 9 is estimated as the mean of the two end boundaries of the class 910.55-990.55 as this class has highest frequency. The class interval containing the largest number of measurements is called the **modal class**. Thus the modal class in Table 9 is "910.55 to 990.55".

For distributions which are symmetrical (that is, where the corresponding frequency polygons or histograms are symmetrical), mean, mode and median coincide. For slightly asymmetric distributions, it has been empirically found that

$$\text{(Mean} - \text{Median)} = \frac{1}{3}\text{(Mean} - \text{Mode)}$$

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

Uses of the Mode

The mode, like the median, can be used as a central location for quantitative as well as qualitative data. If a printing press turns out 5 impressions, which we rate "Very sharp", "Sharp", "Sharp", "Sharp" and "blurred", then the modal value is "Sharp". Like the median the mode is not unduly affected by extreme values. Even if the high values are very high and the low values very low, we choose the most frequent value of the data set to be the modal value. Also, mode can be used even when one or more of the classes are open-ended. However, the mode is not used as often to measure central tendency as are the mean and median. Too often, there is no modal value because the data set contains no values that are repeated more than once. There may be cases when every value occurs the same number of times, or the data sets contain two, three, or many modes. In such cases the mode is not a useful measure.

E15) Classify the following statements as True or False

a) The value of every observation in the data set is taken into account when we calculate its median.

b) Measure of central tendency in a data set refer to the extent to which the observations are scattered.

c) We can compute a mean for any data set, once we are given its frequency

distribution.

d) The mode is always found at the highest point of a graph of a data distribution

A measure of central tendency, as has been mentioned above, gives us a general idea about the average value or the magnitude of the observations. However, two distributions though may have the same mean, say, may differ in respect of several other characteristics. For instance, consider the following data which relate to performances of two suppliers: a manufacturer of a certain electrical equipment purchases 100 cardboard boxes for packing purposes every week from each of two suppliers A and B and the following are the two distributions of defectives in the weekly lots:

Table 12: Distributions of Number of Defective Boxes

Week	No. of Defective Boxes Supplied by	
	A	B
1	12	6
2	3	7
3	7	6
4	11	5
5	0	7
6	2	5
7	9	6
8	1	5
9	1	6
10	14	7
11	2	4
12	10	8
Total	72	72

Observe that both the suppliers have supplied 6 defective boxes on an average per week. Judging from this aspect alone, we may be tempted to conclude that both the suppliers have performed equally. However, a closer look at the distributions reveals that this is not really so. While the number of defectives supplied by A varied widely between 0 and 14 over the weeks, the performance of B has been more or less stable and consistent. The number of defectives supplied by him being more or less around 6. This suggests that the extent of **variation or dispersion** of the given sets of observations from the respective averages together with the averages themselves can give us a much wider scope for comparing the performances.

Similarly, suppose that the average score obtained by students in 10-marks class test is 5 and suppose it is further known that the scores varied between 3 and 6. With such information in view, one can predict his performance with much more confidence than when the scores are known to have varied between 2 and 8.

These examples suggest that the average along with an idea about the scatter or spread of variation or dispersion of the observations about the average give us a more complete picture about the state of affairs than the average alone. The less is the range, the more will be the concentration of observations around the mean. We shall now discuss as to how the dispersion of the observations about the average can be measured.

1.3.2 Measures of Dispersion or Variation

There are two types of summary measures of dispersion; **distance measures** and **measures of average deviation.**

Distance measures describe the variation in the data in terms of the distance between selected measurements. The most frequently used distance measure is the range.

Range: It is defined to be the difference between the largest and the smallest

observations.

In the above table, the range for A is $14 - 0 = 14$ and that of B is $8 - 4 = 4$.

The range, because of its complete dependence on two extreme values, is a quick but not a very accurate measure of dispersion. For instance, both the sets of observations

Set I: 1,1,1,1,1,1,1,1,1,10
Set II: 1, 10,10,10,10,10,10,10

have the same range namely $10 - 1 = 9$; but, evidently, the distributions are very different.

Significant measures of average deviation are **variance** and the **standard deviation**. Both of these tell us an average distance of any observation in the data set from the mean of the distribution.

Mean Deviation and **Standard Deviation**:

We start with considering the case of **ungrouped data**. Let $x_1, x_2, \ldots, x_N$ be N observations and let μ be the mean of these observations. Consider the deviations $(x_j - \mu), j = 1, 2, \ldots, N$ of the observations from the mean. The variation or dispersion is small if the observations $x_1, x_2, \ldots, x_N$ are bunched together in the close neighbourhood of μ and it is large if these are scattered widely away from μ. Thus, higher the dispersion, higher will be the deviations in magnitude. These distances should then be suitably combined to give rise to a consolidated measure of dispersion. However, the arithmetic mean of these deviations, which may initially seem to be a reasonable measure, does not serve our purpose since sum of these deviations is always zero (how?) i.e.

$$\sum_{j=1}^{N}(x_j - \mu) = 0 \tag{5}$$

E16) Verify result (5) for the data 1, 2, 4, 6, 8.

In any case, at this stage, let us appreciate that to measure dispersion, we need to know the extent of absolute deviations of the observations from the mean, and not the direction of these deviations. As such, instead of considering the deviations $(x_j - \mu)$, we should deal with either $|x_j - \mu|$ or $(x_j - \mu)^2$. Following this logic, we define two measures

$$MD(\mu) = \sum_{j=1}^{N} \frac{|x_j - \mu|}{N} \tag{6}$$

$$SD = \sigma = + \left\{ \sum_{j=1}^{N} \frac{(x_j - \mu)^2}{N} \right\}^{\frac{1}{2}} \tag{7}$$

In the above, the measure $MD(\mu)$ is called the **mean deviation** (about the mean); the measure SD denoted by σ is called the **standard deviation**. Both the measures are expressed in the same unit in which the observations are measured. The square of the standard deviation i.e SD^2 is called the **variance** of the data and is denoted by σ^2

For **grouped data** with one variate value for each class, the formula for the mean deviation and the standard deviation are given by

$$MD(\bar{x}) = \left\{ \sum_{j=1}^{k} \frac{|x_j - \mu| f_j}{N} \right\} \tag{8}$$

and

$$SD = \sigma = \left\{ \sum_{j=1}^{k} \frac{(x_j - \mu)^2 f_j}{N} \right\}^{\frac{1}{2}} \quad (9)$$

where $x_1, x_2, \ldots, x_k$ are the distinct observations, f_j is the frequency of x_j and $f_1 + f_2 + \ldots + f_k = N$, the total number of observations.

Formula (9) is used to compute the variance of observations presented in the form of a frequency distribution with k classes with x_j as the class mark of the j-th class, $j = 1, 2, \ldots, k$. As discussed earlier, this will give us only an estimate of the actual variance.

The computation of the variance can be drastically simplified through some simple algebraic treatments as follows:

$$\begin{aligned} \sum_{j=1}^{k}(x_j - \mu)^2 f_j = \sum_{j=1}^{k}(x_j^2 - 2\mu x_j + \mu^2) f_j &= \sum_{j=1}^{k} x_j^2 f_j - 2\mu \sum_{j=1}^{k} x_j f_j + \mu^2 \sum_{j=1}^{k} f_j \\ &= \sum_{j=1}^{k} x_j^2 f_j - 2\mu(N\mu) + \mu^2 N \\ &= \sum_{j=1}^{k} x_j^2 f_j - N\mu^2 \quad (10) \end{aligned}$$

Let us now illustrate these formulas through some examples. You may recall that we computed the mean for the data of Table 5 to be 5.36. Now let us compute the variance and standard deviation of the same data.

Example 10: Consider the data of Table 5 giving the marks of 100 students in a Statistics test in the form given by the following table

Table 13: Calculation of Variance of the Frequency Distribution of marks of 100 Students in Statistics

Score (x_j)	Frequency (f_j)	$x_j f_j$	$x_j^2 f_j$
0	1	0	0
1	2	2	2
2	4	8	16
3	9	27	81
4	15	60	240
5	21	105	525
6	20	120	720
7	15	105	735
8	9	72	576
9	4	36	324
Total	100	535	3219

Thus, the variance using Formula (10) can be obtained as

$$\begin{aligned} \sigma^2 &= (1/100)\{3219 - (100)(5.35)^2\} \\ &= 3.5675 \\ \text{so that SD} &= (3.5675)^{\frac{1}{2}} = 1.8888. \end{aligned}$$

In the same way you can try the following exercise.

E17) Given the following sample of 20 numbers:

12 41 48 58 14 43 50 59 15 45
52 72 18 45 54 78 41 47 56 79

a) Compute the mean, the variance, and the standard deviation.

b) If the largest value in the above set of number is changed to 500 to what degree are the mean and variance affected by the change?

While doing these exercises you must have realised that whenever the x-observations for a given data are large in magnitude, the computations for the calculation of the variance become lengthy. But we can make these computations a bit more manageable Let us see how this can be done.

As before, let the distinct observations be $x_j, j = 1, 2, \ldots, k$ and f_j be the frequency of x_j. Let us write
$y_j = (x_j - a), j = 1, 2, \ldots, k$, where a and b are two pre-specified constants.

Thus, $x_j = a + by_j, j = 1, 2, \ldots, k$.

Hence,

$$\sum_{j=1}^{k} x_j f_j = \sum_{j=1}^{k} (a + by_j) f_j = a \sum_{j=1}^{k} f_j + b \sum_{j=1}^{k} y_j f_j$$

so that, by dividing both sides by N, we note that

$$\mu = a + b\mu' \text{ where } \mu' = \sum_{j=1}^{k} \frac{y_j f_j}{N}.$$

Also, we can write

$$\sum_{j=1}^{k} (x_j - \mu)^2 f_j = \sum_{j=1}^{k} (a + by_j - a - b\mu')^2 f_j = b^2 \sum_{j=1}^{k} (y_j - \mu')^2 f_j$$

so that, we obtain the following result.

Variance of x-observations $= b^2$**(Variance of y-observations).** (11)

This formula is extremely useful from the computational point of view. Basically, the idea is that whenever the x-observations are large in magnitude, we can convert them into y-observations which can be smaller in magnitude if the constants a and b are chosen suitably. We then calculate the variance of the y-observations and simply multiply the same by b^2 to arrive at the variance of x-observations. **Usually a is chosen to be a value in the middle of the range of the observations and b is the class interval.**

Let us now take up an example to see how Result (11) helps us in simplifying the computations for a given set of data.

Example 11: Consider the data of Table 11 as shown in Table 14 on the next page.

The estimated mean μ' of the y-observations is

$$\mu' = \sum_{i=1}^{11} \frac{y_i f_i}{N}$$

$$= \frac{34}{100} = 0.34$$

and the estimated mean μ of the x-observations is

$$\mu = a + b\mu'$$

$$= 950.55 + 80(0.34)$$

$$= 977.75 \text{ hours.}$$

Table 14: Frequency Distribution of Life of 100 Electric Bulbs

Life (hours); class boundaries (exclusive of left boundaries but inclusive of right boundaries)	Frequency f_j	Class marks x_j	$y_j = \frac{((x_j - 950.55)}{80}$	$y_j f_j$	$y_j^2 f_j$
510.55 - 590.55	1	550.55	-5	-5	25
590.55 - 670.55	2	630.55	-4	-8	32
670.55 - 750.55	5	710.55	-3	-15	45
750 - 830.55	8	790.55	-2	-16	32
830.55 - 910.55	13	870.55	-1	-13	13
910.55 - 990.55	26	950.55	0	0	0
990.55 -1070.55	20	1030.55	1	20	1
1070.55 - 1150.55	12	1110.55	2	24	48
1150.55 - 1230.55	6	1190.55	3	18	54
1230.55 - 1310.55	6	1270.55	4	24	96
1310.55 - 1390.55	1	1350.55	5	5	5
Total	100			34	351

Note that here we have taken a=950.55 and b=80.

Using Formula (10), the estimated variance of the **y**-observations is

$$\frac{1}{N}\sum_{j=1}^{11} y_j^2 f_j - \mu'^2 = \frac{351}{100} - (0.34)^2$$

$$= 3.39$$

Now using Formula (11),

the estimated variance of **x**-observations $= \sigma^2$ = b^2(variance of **y**-observations)

= $(80)^2(3.39) = 21696(\text{hours})^2$

$\therefore$ the estimated SD of the **x**-observations $= \sigma$ = $+\sqrt{21696} = 147.2956$ hours.

E18) The following table gives the frequency distribution for the heights (in cms.) of 75 individuals.

Heights (cms.)	Frequency
150.6 - 152.5	6
152.6 - 154.5	7
154.6 - 156.5	9
156.6 - 158.5	13
158.6 - 160.5	17
160.6 - 162.5	8
162.6 - 164.5	9
164.6 - 166.5	6
	75

Find the mean and the s.d. of the above data. Also, estimate the median.

E19) The mean and standard deviation of a set of n observations $x_1, x_2, \ldots x_n$ are $\bar{x}$ and σ_x respectively. The mean and the standard deviation for another set of m observations $y_1, y_2, \ldots y_m$ are $\bar{y}$ and σ_y respectively. Show that the standard deviation of the pooled set $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$ of m+n observations is

given by

$$\sqrt{\frac{n\sigma_x^2 + m\sigma_y^2}{n+m} + \frac{nm}{(n+m)^2}(\bar{x} - \bar{y})^2}$$

We now end this unit by giving the summary of whatever we have covered here.

1.4 SUMMARY

In this unit we have covered the following points

1) To study a given phenomenon/any physical situation, the basic raw material is the **data** (information) relating to it.
2) **Statistics** involves the collection, organisation, analysis and interpretation of data.
3) A collection of data is a **data set** and a single observation from a data set is the **data point**.
4) For a data set to be **useful**, observations need to be organised in order to pick out trends and come to logical conclusions.
5) The number of times any value occurs in a data set is the **frequency** of that value.
6) An organised display of data showing the frequency of each value in a data set against its value or against mutually exclusive classes into which these values fall is a **frequency distribution**.
7) A tabular display of data showing how many observations lie above, or below, certain values is a **cumulative frequency distribution**.
8) **Graphical methods** of presenting data provide an effective way to present data as it gives an idea of the pattern of variation of the random variable (character corresponding to numerical value) at a glance.
9) A **Histogram** is a graph of a data set, composed of a series of rectangles, each proportional in width to the range of values in a class and proportional in height to the number of items falling in the class.
10) A line graph connecting the midpoints of each class in a data set, plotted at a height corresponding to the frequency of the class is a **frequency polygon**.
11) **Ogive** is a graph of a cumulative frequency distribution.
12) Single numbers that describe certain characteristic of a data set are **summary measures**.
13) Summary measures that are commonly used are **measures of central tendency** and **measures of dispersion or variation**.
14) A measure indicating the value to be expected of a typical or **middle data point** is a measure of central tendency. Mean, mode, median are such measures.
15) A measure describing how **scattered** or **spread** out the observations in a data set are is a measure of dispersion. Mean deviation and variance are such measures.
16) **Mean** is a central tendency measure representing the arithmetic average of a set of observations.
17) **Median** is the middle point of a data set, a measure of location that divides the data set into values.
18) The value most often repeated in the data set is the **mode**. It is represented by the highest point in the distribution curve of a data set.
19) In a data set, the average distance of the observations from the mean is the **mean deviation**.

20) In a data set, the average of the square of distance of observations from the mean is the **variance** and its positive square root is the standard deviation.

1.5 SOLUTIONS/ANSWERS

E1) Examples could be taken from situations arising from your daily life experience.

E2) a) population b) sample c) population

E3) b) and d) discrete; a),c) and e) are continuous.

E4) For example data collected to find the average height of females in the age group 15-20 years in Delhi.
Think of other similar examples.

E5) a) 20 and 34, 35 and 49, 50 and 64, 65 and 79, and 80 and 94;

b) 19.5, 34.5, 49.5, 64.5, 79.5, 94.5;

c) 27, 42, 57, 72 and 87;

d) 15.

E6) a) X is a discrete variable; min. score =29, max. score = 97.

b)

Scores	Frequency	Cumulative frequency	
		'less than type'	'more than type'
20-29	1	1	50
30-39	2	3	49
40-49	2	5	47
50-59	3	8	45
60-69	12	20	42
70-79	14	34	30
80-89	12	46	16
90-99	4	50	4

c) 45

d) 29

E7) i) Because of the nature of partition on the range of income, the exact percentage of people earning more than Rs.4500/- cannot be found out; however, an estimate can be tried assuming that the frequency of 152 in the class 4000 - 6000 is uniformly distributed over the interval (4000, 6000). The estimate is obtained by simple linear interpolation as follows:

$$\frac{100}{1000}\left\{\text{Total frequency} - \text{frequency in } (0, 4000) - \frac{152}{6000-4000}(4500-4000)\right\}$$

$$= \frac{100}{1000}\{1000 - 236 - 38\} = 72.6\%$$

ii) As above, an estimate is

$$\frac{100}{1000}\left\{1000 - 40 - \frac{55}{2000-1000}(500)\right\} = 93.25\%$$

iii) $100\{141 + (1000)152/2000\}/1000 = 21.7\%$

E8)

No. of Children	Frequency
0	3
1	9
2	12
3	20
4	8
5	5
6	2
7	1
Total	60

Frequency polygon can be drawn for the above observation as given in Fig.6.

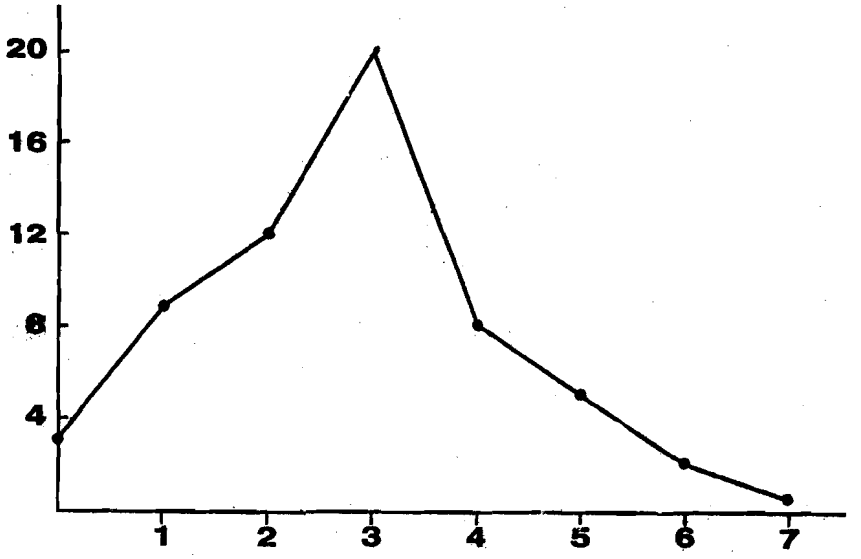

Fig.6

Similarly draw cumulative frequency graph for the above data.

Proportion of families having at least 2 children $= \frac{48}{60} = 80\%$

Proportion of families with not less than 2 children, but not greater than 4 children $= \frac{(12+20+8)}{60} = \frac{40}{60} = 66.67\%$.

E9) Use a histogram.

E10) Use pie diagram

E11) Correct average weekly wage
$= \{80.50(60) - 92.00 + 80.00\}/60 = 80.30$

E12) Suppose there are 100 workers; thus 60 of them are paid @Rs.60/- per day each and 40 of them are paid @ Rs.450/6 = 75/- per day each. The average income then $= \{60(60) + 75(40)\}/100 = 66$/-
After modification, the average $= \{60(1.15)(60) + 40(.80)(75)\}/100 = 65.40$

E13) Mean $= \frac{2909.5}{50} = 58.19$

E14) For the given data $\frac{n}{2} = \frac{70}{2} = 35$
$c = 31, f_m = 16, l = 24.9, L_m = 75$, using Formula (4), median M = 81.23

E15) a) T, b) F, c) F, d) T

E16) $\sum_{i=1}^{5}(x_j - \bar{x}) = -3.2 - 2.2 - 0.2 + 1.8 + 3.8 = 0$

E17) a) $\bar{x} = 46.35, \sigma^2 = 384.56, \sigma = 19.61$

b) Changing the largest value 79, in the given set to 500,
$\bar{x} = 67.40, \sigma^2 = 1484.18$

E18)

Class mark (x)	$y = (x - 157.55)/2$	Freq. (f)	yf	y^2f
151.55	-3	6	-18	54
153.55	-2	7	-14	28
155.55	-1	9	-9	9
157.55	0	13	0	0
159.55	1	17	17	17
161.55	2	8	16	32
163.55	3	9	27	81
165.55	4	6	24	96
Total		75	43	317

Thus,

$$\bar{x} = 157.55 + 2\bar{y} = 157.55 + 2(43/75) = 158.70;$$

$$\sigma_x = 2, \sigma_y = 2\sqrt{\left(317 - \frac{(43)^2}{75}\right)/75} = 7.80$$

$$M = 158.55 + \{(0.5 - 35/75)(75)(2)\}/75 = 158.62$$

E19) The pooled set has mean $\bar{u}$ and variance σ^2, say. Then

$$\bar{u} = \frac{n\bar{x} + m\bar{y}}{n + m}$$

$$\begin{aligned}
\text{Also, } (n + m)\sigma^2 &= \sum_{j=1}^{n}(x_j - \bar{u})^2 + \sum_{j=1}^{m}(y_j - \bar{u})^2 \\
&= \sum_{j=1}^{n} x_j^2 + \sum_{j=1}^{m} y_j^2 - (n + m)\bar{u}^2 \\
&= \left(n\sigma_x^2 + n\bar{x}^2\right) + \left(m\sigma_y^2 + m\bar{y}^2\right) - (n + m)\bar{u}^2 \\
&= \left(n\sigma_x^2 + m\sigma_y^2\right) + \frac{nm}{n + m}(\bar{x} - \bar{y})^2
\end{aligned}$$