UNIT 11 STATISTICS

Structure

11.1 INTRODUCTION

- The average height of Indian women is less than that of European women.
- Kerala is the most literate of the Indian states.
- Smokers are more likely to get lung cancer than non-smokers.

You must have read similar statements in your daily newspaper. Such statements are made with the help of statistics on the basis of the information gathered in surveys. The word statistics is quite old but its meaning has undergone a good deal of change. It is derived from the Greek word statistik for the study of facts and figures relating to politics. The root of statistik is the Latin status, meaning state. Thus, statistics originally meant information, useful for the affairs of the state, such as raising an army or collecting taxes. Raising an army required a head-count of the citizens (a primitive form of census), and to collect the taxes the state needed detailed information about their economic status. However, in the last two centuries or so, the word statistics has acquired a different connotation.

Statistics now means a science – the science which provides the techniques and methodology for collection, processing, presentation and analysis of numerical data. In this unit, we'll explain the meanings of various statistical terms. We shall be using these terms frequently in this and the next block. We shall also discuss the methods of summarisation of data on one variable. You will also find the procedures for calculating important characteristics of variation of statistical data in this unit.

Objectives

After reading this unit, you should be able to

- define various statistical terms like statistical data, population etc
- obtain a frequency distribution from the raw data on a variable
- calculate mean, median and mode from ungrouped and grouped data
- calculate standard deviation, mean deviation and semi-interquartile range from ungrouped and grouped data.

11.2 BASIC DEFINITIONS

Suppose a dam is to be built over a river, and architects have been appointed to design this dam. They will have to consider a lot of factors before they decide on the exact design. For example, among other things, they will need to know the maximum possible rainfall in that region in a day. Thus, before drawing up plans for the dam, it will be necessary to look at the daily rainfall figures in the region which have been collected over the years. But of course, a mere look at the collection of figures will not be enough ! These will have to be organised in tables and then analysed. Now this is what statistics is all about! We have already mentioned in the Introduction, that statistics provides techniques for collecting, processing and analysing numerical data, like the rainfall figures in the above example.

We shall now start our study of the various aspects of statistics by studying the definitions of some terms which are commonly used in it.

11.2.1 Statistical Data

Statistics deals with data which show irregular or unpredictable variation, and which may arise from any human endeavour or natural phenomenon. Such data are called **statistical data**. We have already seen an example of such data in the rainfall figures mentioned in the beginning of this section.

Thus, statistical data possess two important characteristics:

i) They are numerical, and
ii) they show fluctuations which cannot be fully explained or predicted.

Usually, but not always, the data are large-scale. They are usually obtained through some systematic process of measurement or counting. In the former case, they have an appropriate unit (gms, cms, rads and so on), while in the latter case they are numbers.

For example, for the data on rainfall figures which we have been talking about, the unit will be cms. Sometimes we may also need to collect data about the number of members in each family in a Madras suburb, say. In this case the data consists of just numbers.

Numerical data obtained through observation or experimentation are usually statistical. This is because the results of the experiment vary, more or less unpredictably, from place to place or time to time or situation to situation. You can see the same tendency in the data coming from such diverse sources as trade, commerce, business, industry, finance and government, or those obtained by experimental scientists, social scientists, engineers and technologists. In fact, variation can occur even when the conditions of observation or experimentation are sufficiently controlled. Thus, when a physicist or a chemist, working in the very controlled environment of a laboratory, repeats an experiment several times, he may find that no two of his results tally exactly. A biologist similarly may find that identical twins brought up in the same family do not show identical physical or mental traits.

In actual practice social scientists, biologists, engineers and technologists usually do not work in such highly controlled environments. Naturally, in such cases the variations become wider. The science of statistics is concerned with the study of all aspects of such variation. Now, having defined statistical data, let us discuss the meanings of some other terms in the next sub-section.

11.2.2 Random Variables

You have already become familiar with variables in Block 1. A variable is a quantity which can take on any value from a given set, called its domain. We'll denote a variable by a capital letter like X, Y, Z. For example,

i) Suppose we are studying the daily rainfall at Bombay. We can take the daily rainfall to be the variable X (in mm). Then X can take any value from 0 to

300, (300 mm being the highest rainfall ever recorded).Thus, X is a variable with a domain which consists of all real numbers from 0 to 300, that is, the interval [0,300].

ii) Suppose we are interested in the number X of petals in a rose belonging to a particular variety. Here X is a variable on the domain given by the set {5, 10, 15, ..., 60}, since the number of petals in a rose is a multiple of 5 and 60 is the maximum number observed for this variety.

iii) If we are studying the number of students (X) admitted to the post-graduate Mathematics course in a year in Calicut University Centre, what will be the range of the variable X? X will vary from 0 to 20, where 20 is the maximum number of students admitted in a year.

Domain and **range** are used interchangeably to denote the span of values that a variable can take.

Now after these examples of variables let us see what is meant by a random variable.

Basically, if an experiment is performed, and some quantitative variable, denoted by X, is measured or observed, then the variable X is called a **random variable,** since the values that X may assume in the given experiment depend on chance.

We can also say that a random variable is a numerical quantity, the value of which is determined by an experiment. In other words, its value is determined by chance. Thus, the variables discussed in the three examples in the beginning of this subsection are all random variables. The various values that a random variable takes is called statistical data. Thus, we can say that the set of values of the random variable, which are obtained by a process of measurement or observation or counting carried out on each individual of the group under study, forms the **statistical data** on the random variable.

Now let us take a close look at these examples again.

In example ii) the random variable can take any one of the **12** values, 5, 10, 15, ..., 60.

In example iii) the random variable can take any one of the **21** values 0, 1, 2, ..., 20.

In these two cases you will see that the random variable takes isolated values.

What about example i) ? Here X can take any value between 0 and 300. The rainfall in Bombay on any particular day can be 5.65 mm, 72.8 mm or even 211.0282 mm (provided we can measure it that accurately!). This means, X can take the value 5.65 or 72.8 or even 211.0282. Thus, you will see that in this case X does not take isolated values but its domain is the entire interval [0,300]. Variables of this type are called continuous random variables, and variables of the type discussed in examples ii) and iii) are called discrete random variables.

Thus, whenever all the possible values that a random variable can take can be listed (or counted), then the random variable is said to be **discrete.** On the other hand, a random variable, which can assume any value in one or more intervals is called a **continuous random variable**.

Let us consider the random variable where X = height of a plant (in cm.). You would agree that X is a continuous random variable. In fact, when we say that X=6.8, it does not mean that the height of the plant is exactly 6.8 cms. It only means that heights are measured correct to one place of decimal and the plant could have any height ranging from 6.75 cm to just below 6.85 cm.

Some other examples of a continuous variable are:

i) the waiting time at a bus stop,
ii) the weight of a new-born baby,
iii) the percentage of calcium in a sample of water.

Look at the data given in Table 1 and Table 2. The random variable X in Table 1 is continuous, but that in Table 2 is discrete.

Table 1: Values of X = Systolic Blood Pressure (in mm) of 70 Adult Males

151.7	123.4	120.6	123.8	96.3	110.9	115.7
117.6	112.3	127.9	120.0	122.0	121.9	117.0
123.3	110.4	136.6	108.6	124.6	146.8	110.6
115.7	155.5	130.1	161.8	116.2	121.1	120.8
117.6	116.4	99.0	137.2	133.7	127.9	121.7
131.3	110.4	109.6	138.5	120.6	143.7	163.1
186.3	145.1	130.5	148.5	105.7	116.0	153.7
168.4	123.2	118.6	133.9	136.7	120.0	138.6
113.8	107.3	118.8	152.0	153.5	177.8	125.3
132.1	144.2	123.6	117.5	128.0	101.5	147.7

Table 2: Values of X = Number of seeds in a packet which germinated in a fortnight when 80 packets of 100 seeds each were planted

86	95	92	89	92	92	91	88	88	94
90	88	93	92	89	90	86	88	93	92
91	87	91	94	92	94	92	92	91	92
88	94	93	91	92	91	92	90	89	86
93	91	89	90	90	89	84	89	88	93
92	95	86	93	93	93	94	91	90	89
89	95	92	94	94	90	86	93	88	93
86	94	90	87	93	91	90	92	93	94

Actually, you may say that the random variable in Table 1 also looks discrete. But as we have noted in our earlier example of heights of plants, this random variable appears discrete only because its values are given accurately to a specific decimal place. So the discreteness is apparent, not real.

See if you can do this exercise now.

E1) The random variables below were recorded for the animals in the Calcutta zoo. Which of these are continuous, and which are discrete ?

a) age
b) year of birth
c) height
d) number of animals admitted in a calendar year
e) weight

Before proceeding to the next sub-section, we would like to tell you that there are some random variables which are neither discrete nor continuous. But in this course we shall be concerned only with either discrete or continuous random variables.

11.2.3 Individual, Population and Sample

Consider the set of data given in Table 1. They relate to the values of a continuous random variable X = Systolic blood pressure (in mm). Now systolic blood pressure is a variable characteristic possessed by each human being. A group of 70 adult males was examined for this characteristic, and the results of this examination are given in Table 1.

Table 2 shows the values of a discrete variable X = the number of seeds which germinated in a fortnight when the contents of a packet of 100 seeds were planted. To obtain the values of X, a specific group of 80 packets was used, and the seeds which germinated from each packet were counted.

So **individuals** in statistics are simply distinct entities which possess some variable characteristic, and hence yield the values of a random variable.

Thus, in Table 1, each member of the group of 70 adult males is an individual and in Table 2, each of the 80 packets of seeds is an individual.

E2) Can you identify the individual and the random variable in the following cases ?

a) A study of the variations in the sizes of the families resident in a locality.
b) A survey of the percentage of alcohol in wines of 20 different brands.
c) A study regarding the volume of sales of TV sets of a particular company from year to year.

We now come to the concepts of population and sample. By definition, a **population (or universe)** consists of all the individuals one is interested in, and a **sample** is a part of the population. In other words, population is the aggregate or the totality of individuals about whom we wish to draw conclusions, and a sample is a part of the population.

Looking at Table 1 again, let us ask the question: "Why were the data collected?"

The obvious answer is to gain some idea about the pattern of variation of systolic blood pressure and to study the important features of the variation. In fact, the study of the systolic blood pressure of these 70 individuals may help us to draw conclusions about the systolic blood pressure variations in all adult males. In this case, the 70 individuals investigated form a sample, and the population consists of all adult males.

In most cases you will find that it is practically impossible to collect data about the entire population, and hence we have to draw inferences from the study of a sample of the population. In fact, many experimental scientists have to work with sample data only, and sometimes the amount of data they can collect is quite small. The problem of deriving valid conclusions from data of small samples is also tackled through statistical methods. We shall discuss these things in detail in Unit 15.

Assuming that the data have been collected, let us now turn our attention to the organisation of the data in a meaningful way.

11.3 FREQUENCY DISTRIBUTIONS

When we have data on a random variable from each individual in a population, a major problem that is usually faced is that of bulk. In most cases the population is so large, that from the raw data on the variable, it is impossible to form any idea about the pattern and characteristics of its variation. For example, the pattern of the variation of systolic blood pressure is not clear from the data in Table 1. The data in Table 2, similarly, does not give us any clear indication of the pattern of variation in the number of germinated seeds. It is difficult to answer even simple questions like

"Between which two values does the random variable vary",
"What value is assumed most frequently", and so on.

So it is necessary to condense or summarise the data in such a manner that information which is of interest to us is retained and highlighted, and irrelevent information is eliminated. This is best done by classifying or grouping the individuals. Let's see how this is done. We shall first consider the problem of classifying a discrete random variable.

11.3.1 Discrete Random Variables' Case

If the random variable is discrete and takes only a few different values, we can count the number of times each value is attained. This number is called the frequency of that value of the random variable. We follow the "tally marking" procedure to obtain the frequencies: We first write down the distinct values taken by the variable in an increasing or a decreasing sequence (see the first column in Table 3, where we have classified the raw data given in Table 2). Then we go over the raw data only once, but carefully. Each individual is placed in its own class by putting a tally mark (|) against its value. Every fifth tally against a particular value is placed horizontally on the previous four (𝍸), so that the tallies are automatically divided into groups of five (see the second column in Table 3). When the process is complete, the number of tallies against each value can be easily counted. The count against each value of the random variable gives the frequency of that value. The sum total of the frequencies of all the values is called the total frequency. Table 3 below shows the frequency distribution corresponding to Table 2.

Table 3: Frequency Distribution of the data in Table 2

No. of Seeds	Tally Marks	frequency
84	𝍷	1
85		0
86	𝍸 𝍷	6
87	𝍷𝍷	2
88	𝍸 𝍷𝍷	7
89	𝍸 𝍷𝍷𝍷	8
90	𝍸 𝍷𝍷𝍷𝍷	9
91	𝍸 𝍸	10
92	𝍸 𝍸 𝍷𝍷𝍷𝍷	14
93	𝍸 𝍸 𝍷	11
94	𝍸 𝍷𝍷𝍷𝍷	9
95	𝍷𝍷𝍷	3
		80

You can now apply this procedure to do this exercise.

E3) The number of peas X in each of 150 pea-pods is given in the table below.

4	3	5	4	4	5	3	4	5	5	2	3	4	4	5.
5	3	4	2	2	2	2	4	4	3	3	3	3	3	4
4	6	6	4	2	3	3	4	3	3	2	4	3	3	3
3	2	4	3	5	3	3	3	2	3	4	4	3	2	2
2	3	5	3	4	4	2	2	3	2	5	4	2	3	3
1	6	4	3	4	3	2	2	3	3	4	2	3	3	2
7	3	4	3	3	4	3	5	3	2	2	4	2	5	3
6	2	4	3	4	3	4	2	3	4	4	3	4	5	3
4	3	3	5	4	6	4	4	4	5	3	4	4	3	3
4	5	5	2	2	4	3	3	6	4	4	4	3	1	4

The largest value of the variable is 7, and the smallest is 1. Obtain its frequency distribution.

You must have noticed that there has been a good deal of summarisation by classifying the raw data in a frequency distribution as in Table 3 and E3). Thus, after doing E3) you will find that instead of having to tackle 150 figures, we now have to tackle only 7. We can also make some observations about the pattern of variation of the random variable. For example,

i) 16 out of the 150 individuals have the value 5.
ii) 98 out of 150 individuals have either the value 3 or the value 4
iii) 3 has the highest frequency, that is, it is the most frequently occurring value.

The pattern of variation of the random variable has also become somewhat clearer. The only information that has been lost is, "What value does a particular individual take?" But this information is not of any interest in statistical studies, which are concerned with the whole aggregate of the individuals, and not with specific individuals.

The procedure which we have discussed is useful when a discrete random variable takes only a few distinct values. But when a discrete random variable takes a large number of values, it may not be worthwhile to classify the individuals according to single values of the variable. We can club together several consecutive values in a class, and obtain the frequencies of these classes. The following example illustrates this.

Example 1: Consider the data of Table 2 again. We have seen that the random variable, the number of seeds which germinate out of 100 seeds in a packet, takes values from 84 to 95. In Table 3 we have formed a frequency distribution by taking each individual value. Here we shall club together two values of the random variable to obtain six classes, instead of twelve. We can then obtain the frequency distribution by tally marking as before.

Table 4

Number of seeds germinating out of 100 seeds in a packet	Tally Marks	Frequency
84-85	|	1
86-87	𝍸 |||	8
88-89	𝍸 𝍸 𝍸	15
90-91	𝍸 𝍸 𝍸 ||||	19
92-93	𝍸 𝍸 𝍸 𝍸 𝍸	25
94-95	𝍸 𝍸 ||	12
		80

Note that now some information of real statistical interest has been lost. We can no longer answer questions like: How many individuals had a value of 88? What fraction of the individuals had values between 85 and 88 (both inclusive)?

However, if it is assumed that the individuals in any class are distributed uniformly, we can roughly estimate the required numbers. Thus, on the above assumption, 15/2 individuals have the value 88, and 1/2 an individual has the value of 85.Thus the approximate number of individuals who have values between 85 and 88 is $\frac{1}{2} + 8 + \frac{15}{2} = 16$.

Here is an exercise for you.

E4) In an objective type examination, there are 100 question items. An item is scored 1 if the answer is correct, otherwise it is scored 0. The total scores X received by 100 students are given in the table below.

75	86	66	86	50	78	66	79	68	60
80	83	87	79	80	77	81	92	57	52
58	82	73	95	66	60	84	80	79	63
80	88	58	84	96	87	72	65	79	80
86	68	76	41	80	40	63	90	83	94
74	66	74	76	68	82	59	75	35	34
65	63	85	89	79	77	76	74	76	78
75	60	96	74	73	87	52	98	88	64
76	69	60	74	72	76	57	64	67	58
72	80	72	56	73	82	78	45	75	56

Answer the following questions on the basis of the data given in the table.

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

b) Using the classes 30 – 39, 40 – 49, ..., 90 – 99. draw up the frequency distribution of X.

c) What proportion of the students score above the pass mark of 50?

d) If a score of 80 or more fetches the A grade, what proportion of students score A in the examination?

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

f) Can you say what is the approximate number of students who score between 55 and 74, using only the frequency distribution?

So far we have seen how to obtain the frequency distribution of a discrete random variable. In the next sub-section we shall take up the case of continuous random variables.

11.3.2 Continuous Random Variables' Case

We follow a slightly different approach to draw up a frequency distribution of a continuous random variable. We have seen that a continuous random variable can take any value in a given interval. For example, the height of a particular variety of mango tree could be anywhere between 1 m and 10 m. Now since the number of different values which a continuous random variable can take is infinite, counting the frequencies of each value, or of several isolated values (as in Table 3 and Table 4) is not feasible. However, in this case we can break up the domain into non-overlapping sub-intervals. These sub-intervals are called **class intervals**. We then find out the frequency of each class interval by counting how many of the individuals have their values in that interval. We can, as before, use the tally-marking procedure for this purpose. The following example will make this discussion clear.

If the blood pressure of any individual lies between 186.25 and 186.35 we record it as 186.3 Similarly, if it is between 96.25 and 96.35 we record it as 96.3. That is why the range = 90.1 units.

Example 2: Consider the data of Table 1 on systolic blood pressure of adult males. The minimum and maximum values of the random variables are 96.3 and 186.3. The range of variation of the variable, therefore, is 186.35 - 96.25 = 90.1 units (see margin remark). Now this is a rather inconvenient figure, because it cannot be divided by any suitable integer. So we extend the range by assuming that the minimum is 92 and the maximum is 191.9. This is always permissible.

The range is now 191.95 – 91.95 = 100 units, and we can have ten class intervals given by 92 – 101.9, 102 – 111.9, 112 – 121.9, 122 – 131.9, 132 – 141.9, 142 – 151.9, 152 – 161.9, 162 – 171.9, 172 – 181.9, and 182 – 191.9. By using the tally marking procedure, we obtain the frequency distribution of the random variable (see Table 5)

Table 5: Frequency Distribution of Systolic Blood Pressure (in mm) of 70 Adult Males

Systolic Blood Pressure (mm)	Tally Marks	Frequency
92 – 101.9	|||	3
102 – 111.9	𝍸 |||	8
112 – 121.9	𝍸 𝍸 𝍸 𝍸 |	21
122 – 131.9	𝍸 𝍸 ||||	14
132 – 141.9	𝍸 |||	8
142 – 151.9	𝍸 ||	7
152 – 161.9	𝍸	5
162 – 171.9	||	2
172 – 181.9	|	1
182 – 191.9	|	1
Total		70

You should remember the following points while drawing up the frequency distribution of a continuous random variable.

i) The recorded value of the variable is always approximate, correct to a specific place of decimal.

Thus, if somebody's blood pressure is recorded as 151.7 mm., it can really be anything from 151.65 to just below 151.75. In general, the range is given by maximum value – minimum value + 1 unit.

ii) The number of class intervals to be used depends on the size of the data. Further, if we are ready to condense the data to a great extent, and are prepared to sacrifice the details, we can take only a few class intervals. But, if we want detailed information from the data, then we'll have to take a large number of class – intervals.

If we have too many class intervals with comparatively few individuals, then the regularity of the frequencies in different class intervals is lost. This regularity has to be maintained, otherwise the pattern of variation of the random variable becomes obscure. Moreover, the more the number of class intervals, the less is the amount of summarisation.

But, on the other hand, if the number of class intervals is too small, then quite a good deal of information is lost. From the frequency distribution in Example 2, we can say what fraction of individuals have blood pressures between, say 122 and 131.9 mm. But we cannot determine the fraction of individuals whose blood pressure is between 122 and 125.9 mm. Weighing both these factors, we usually have between 8 to 20 class intervals, depending on the size of the data and the range of variation.

iii) The class intervals need not be equal; but for the sake of convenience, we frequently make the class intervals equal in length. For this purpose, it is quite permissible to extend the range by reducing the minimum and increasing the maximum. The extension should be, as far as possible, equal at the two ends.

iv) If there is some specific reason, we may need to have class intervals of different length. For example, since mortality is very heavy in the first few days of life, in a study of infant mortality one may draw up a frequency distribution of the age at death with very unequal class intervals. This is illustrated in Table 6.

Table 6: Age Distribution of Deaths of Infants in a country in a year

Age at Death	Frequency
under one day	26,665
1 day	8,364
2 days	6,344
3-6 days	12,375
1 week	10,911
2 weeks	7,717
3 weeks but below one month	6,212
1 to less than 2 months	15,362
2 to less than 3 months	12,066
3 to less than 6 months	27,487
6 to less than 9 months	20,409
9 months to less than a year	17,112
Total	1,71,024

Table 7

Income (in rupees)	Frequency
Under 200	36,857
201 – 299	22,374
300 – 399	19,408
400 – 499	15,049
500 – 599	9,529
600 – 699	6,833
700 – 799	3,950
800 – 899	2,785
900 – 999	5,520
1000 – 1499	2,197
1500 – 1999	2,197
2000 – 2499	1,027
2500 – 2999	579
3000 – 4999	847
5000 and over	523
Total	1,29,663

v) If the range is wide because a small fraction of the individuals have very low values and/or very high values, and the rest have intermediate values, a fine division may be required for these extreme values only. In this case we can keep the first class interval and/or the last class interval open. This type of situation occurs verv often when a random variable like income is considered. See Table 7

An open first interval means there is no lower limit for that class.
An open last class interval means the last class does not have any upper limit.

If you remember these five points, you should not have any difficulty in obtaining the frequency distribution of any continuous variable. We'll now define some terms which are often used in connection with frequency distributions.

The end-points of a class interval are called **class limits**. For example, in Table 5, the class limits are 92, 101.9, 102, 111.9. etc. Further, for the class 92 – 101.9, 92 is called the **lower class limit**, and 101.9 is called the **upper class limit**.

In all the frequency distributions considered so far, you will find that there is a gap between the upper limit of one class, and the lower limit of the next class. In Table 5, the upper class limit of the first class is 101.9 mm, and the lower class limit of the next class is 102 mm. Thus, there is a gap of 0.1 mm between the two. We divide the difference in such a gap by two, and add it to the concerned upper class limit. We subtract the same amount from the concerned lower class limit. This gives us the real **boundary** between these classes. Therefore, for the first and the second class in Table 5, we add $\frac{0.1}{2}$ to 101.9, and subtract $\frac{0.1}{2}$ from 102 to get 101.95. This is the real boundary between these two classes. Similarly, the other boundaries in Table 5 are 111.95, 121.95 and so on.

The difference between the upper and lower boundaries of a class is called the **length of the class interval**. What is the length of the class intervals in Table 5? 10, of course.

The mid-point of a class interval is called its **class value**.

Thus, the class value of the first class in Table 5 is $\frac{92+101.9}{2}$ = 96.95, that ot the second class is 106.95, and so on.

After writing the frequency distribution corresponding to the given data, we can also represent it pictorially. Such representations are very useful, because through them we can get a good idea of the pattern of variation of the random variable at a glance. We shall take up the diagrammatic representation of frequency distributions in the next sub-section.

11.3.3 Diagrammatic Representation

We normally use a frequency polygon or a histogram to diagrammatically represent a frequency distribution.

In a **frequency polygon** the frequency of a class interval is plotted against its class value. The plotted points are then joined together by straight lines. It is thus tacitly assumed that the frequency of a class interval is concentrated at its mid-point. The frequency polygon may be used when the random variable is discrete, particularly when each value of the variable corresponds to a class. In Fig. 1 (a) and (b) you can see the frequency polygons corresponding to the data in Table 3 and Table 5, respectively.

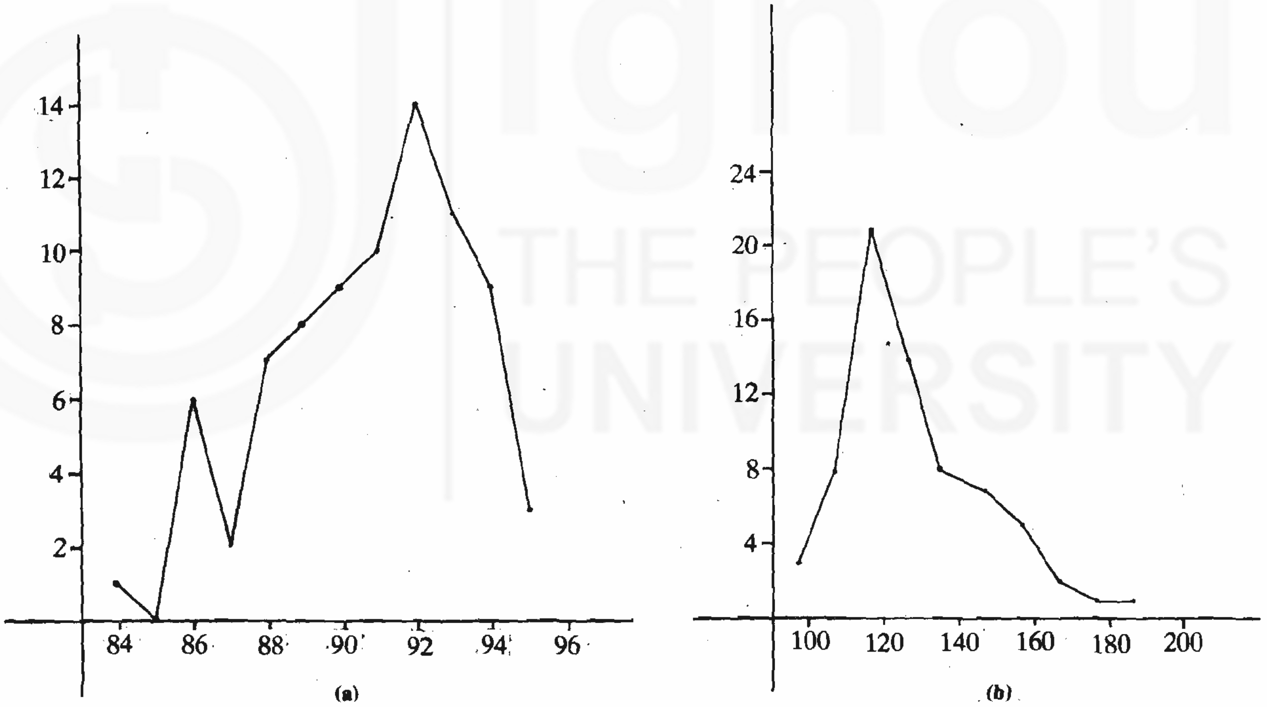


Fig. 1

For plotting a histogram corresponding to a given frequency distribution, we plot class boundaries on the x-axis. Then on the y-axis we plot **frequency densities**. But what is frequency density? It is the frequency per unit length of the class interval. So, if the frequency of the class 40-49 is 54, then its frequency density is

$\frac{54}{49.5-39.5} = 5.4$ (where 39.5 and 49.5 are the class boundaries). Now going back to the histogram, we plot class boundaries on the x-axis, and frequency densities on the y-axis. Then we draw rectangles with class intervals as bases, and the corresponding frequency densities as heights. 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 areas of these rectangles represent the frequencies of the corresponding classes.

You may be wondering why we plot frequency densities on the y-axis, and not frequencies themselves. In fact, if the class intervals are all equal, we do plot frequencies on the y-axis. But we have to plot frequency densities in those cases where the classes are of unequal lengths to ensure that the areas of the rectangles in the histogram correspond to the frequencies of the corresponding classes.

Fig. 2 shows a frequency polygon and a histogram representing the frequency distribution of Table 8.

Table 8 : Head Length (in mm) of a Group of Prisoners, Age 25 – 30

Class limits	Class Boundaries	Class value	Frequency
172–175	171.5–175.5	173.5	3
176–179	175.5–179.5	177.5	9
180–183	180.5–183.5	181.5	29
184–187	183.5–187.5	185.5	76
188–191	187.5–191.5	189.5	104
192–195	191.5–195.5	193.5	110
196–199	195.5–199.5	197.5	88
200–203	199.5–203.5	201.5	30
204–207	203.5–207.5	205.5	6
208–211	207.5–210.5	209.5	4
212–215	211.5–215.5	213.5	2
216–219	215.5–219.5	217.5	1

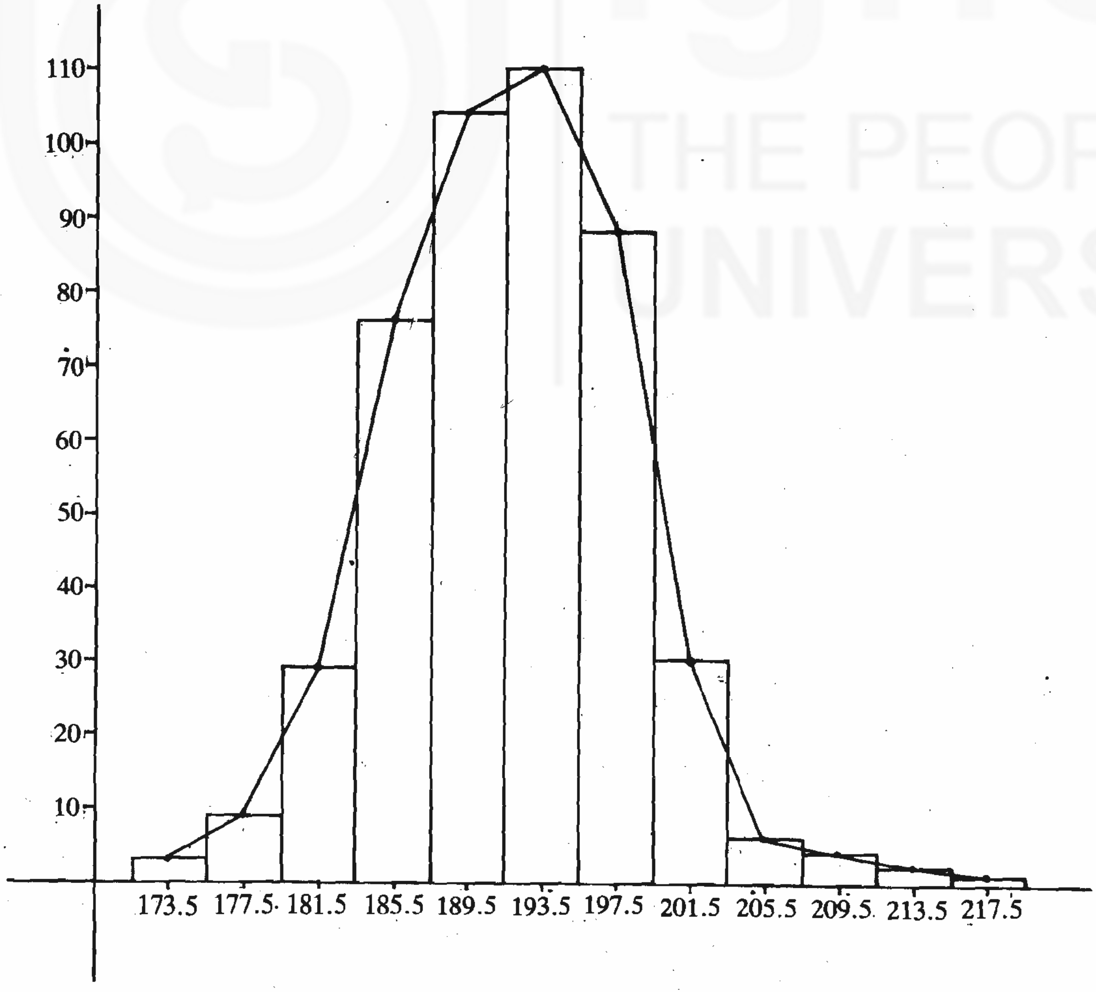


Fig. 2

We are sure you will be able to do this exercise now.

E5) Draw the frequency polygon and the histogram of the frequency distribution given in the following table.

Age	Frequency
0 – 4	270
5 – 9	425
10 – 14	616
15 – 19	939
20 – 24	881
25 – 29	579
30 – 34	381
35 – 44	591
45 – 54	456
55 – 64	275
65 – 74	141
75 & above	45

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. However, often we have to condense the data still further, particularly for purposes of comparison. For example, suppose we want to compare two varieties of paddy plants with respect to their heights. Let us assume that their frequency distributions are available to us. But we cannot carry out any comparisons unless more sophisticated statistical methods are used. Usually it is sufficient to know only two things:

*whether the typical heights of the plants of the two varieties differ, and
**whether one variety shows more fluctuation in height than another.

In general, these two characteristics of a random variable, namely,

i) the typical value or average, and
ii) the variability or scatter, are the two most important features of its variation. It is necessary, therefore, to define some quantitative measures of these characteristics and the methods of their computation from statistical data. In the next two sections we shall discuss the methods of calculation of these measures both for ungrouped and grouped data. Let's talk about the average first.

11.4 MEASURES OF CENTRAL TENDENCY

Very often we find that the values of a random variable, in spite of their variation, show a distinct tendency to cluster around a typical central value. A frequency distribution or histogram makes this tendency clear. Usually, at the beginning of a frequency table, the frequencies are small. They gradually rise to a maximum near about the middle of the table, and then taper off to low values again towards the end. You can check this from the tables given in the unit, and from Fig. 1 and Fig. 2. There may be some irregularities in this pattern; but mostly these are due to insufficient data or to the fineness of the class intervals.

We can look at the centrally located value towards which other values have a tendency to concentrate, as typical or representative of the whole population. It is therefore called a **measure of central tendency** or simply an **average** value of the variable.

There are several measures of central tendency or averages in statistics. Here we shall discuss three of these: the mean, the median and the mode. We should, in any given situation, use the most appropriate one, keeping in mind its status as the representative value of the aggregate.

11.4.1 Arithmetic Mean

The most commonly used measure of central tendency is the arithmetic mean, or simply the mean, of the random variable. It is denoted by the greek letter μ or by M. To keep track of the random variable, if necessary, we use the symbols μ_x or M_x. Now let us see how to calculate the mean from ungrouped data.

Calculation of Mean from Ungrouped Data

For ungrouped data on a random variable, the mean μ is obtained by summing all the values and dividing the sum by the total number of values. Thus, the mean μ_x of the random variable X, whose values are given by $x_1 = 30$, $x_2 = 40$, $x_3 = 60$ and $x_4 = 80$, is

$$\mu_x = \frac{30+40+60+80}{4} = 52.5.$$

Algebraically if x_1, x_2, ..., x_N are the values of a random variable X, then

$$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \text{...(1)}$$

Now if we construct a new random variable U by subtracting a constant A from each value of X and dividing the result by another constant C, that is, if we take

$$u_i = \frac{x_i - A}{C}, \; i = 1, 2, \ldots\ldots, N,$$

then the arithmetic means of X and U are connected by the formula,

$$\mu_x = C\,\mu_u + A \qquad \text{...(2)}$$

A is called the base and C the scale, and the process described above is called "a change of base and scale of X"

This process can save a lot of the labour in calculating the mean. Here is an illustration.

If $x_1 = 25$, $x_2 = 35$, $x_3 = 65$ and $x_4 = 85$,

then $\mu_x = \dfrac{25+35+65+85}{4} = 52.5$

We can also calculate μ_x by change of base and scale.

We can choose A = 45 and C = 10. Then

$$u_1 = \frac{25-45}{10} = -2,\; u_2 = -1,\; u_3 = 2 \text{ and } u_4 = 4.$$

$$\therefore\; \mu_u = \frac{-2-1+1+4}{4} = \frac{3}{4} = 0.75$$

$$\mu_x = A + C\mu_u$$

$$= 45 + 10 \times 0.75 = 52.5$$

You can see that if we change the base and scale suitably, we have to work with smaller numbers (and there are lesser chances of making mistakes !).

Try to do this exercise now.

E6) For the data x_1, x_2, x_3, ... , x_N, show that

$$\sum_{i=1}^{N} (x_i - \mu) = 0 \text{ where } \mu \text{ is the mean of the } x_i\text{'s.}$$

Now let us see how to calculate the mean from a grouped data.

Calculation of Mean from Grouped Data

To calculate the mean of a random variable from a frequency distribution we first obtain the **class values**, which are the mid-points of the class intervals. The class values are then multiplied by the corresponding class frequencies, and the products are added. The sum, divided by the total frequency, gives an approximate value of the mean. In fact, the exact value of the mean can only be obtained by using the raw data or by using individual values of the random variable as classes. However, the difference between the exact value of the mean and its approximate value obtained from a frequency distribution will not be large if the length of the class intervals is small compared to the range. The procedure for calculating the mean of a grouped data should be clear from the following examples.

Example 3: The number of heart-beats per minute of 20 persons admitted to a hospital are given below.

78, 77, 79, 80, 75, 81, 77, 78, 82, 79, 78, 76, 80, 78, 77, 78, 76, 79, 77, 78.

Calculate the mean of this data from the frequency distribution with class intervals of length 2, and compare this with the exact value of the mean.

Solution : The frequency distribution of the data, with four class intervals of length 2 is

No. of heart beats per minute	Frequency
75 – 76	3
77 – 78	10
79 – 80	5
81 – 82	2
Total	20

The class-values and the calculations necessary to obtain the mean are:

Class values	Frequency	Class value × Frequency
75.5	3	226.5
77.5	10	775.0
79.5	5	397.5
81.5	2	163.0
Total	20	1,562.0

$$\text{Mean } \mu = \frac{1562}{20} = 78.10$$

The mean, calculated from the raw data, is,

$$\mu = \frac{1563}{20} = 78.15$$

Thus, the exact value of the mean differs by only 0.05 from that obtained from a frequency distribution with 4 class intervals.

Example 4 : Compute the mean of the data in Table 4.

Solution:

No. of Germinating seeds	Class values	Frequency	Class value × frequency
84 – 85	84.5	1	84.5
86 – 87	86.5	8	692.0
88 – 89	88.5	15	1327.5
90 – 91	90.5	19	1719.5
92 – 93	92.5	25	2312.5
94 – 95	94.5	12	1134.0
Total		80	7270.0

$$\text{Mean} = \frac{7270}{80} = 90.875$$

In symbols, if $x_1, x_2, \ldots\ldots, x_k$ are the class values and the corresponding frequencies are $f_1, f_2, \ldots\ldots, f_k$, then,

$$\mu_x = \frac{\sum_{i=1}^{k} x_i f_i,}{N}, \text{where N is the total frequency.}$$

Short Method of Calculating the Mean

Suppose we take $u_i = \frac{(x_i - A)}{C}$ where A is one of the class values, and C is the length of the class intervals. Then, we know that the relation between μ_x and μ_u is given by

$$\mu_x = C\,\mu_u + A$$

This result makes it much easier to calculate the mean of a frequency distribution with class intervals of the same length as you can see from the following example.

Example 5: Compute the mean from Table 5 using the short method.

Solution: Taking A = 136.95 and C = 10, the length of the class intervals, we get the following table.

Systolic Blood Pressure (in mm)	Frequency (f_i)	Class values (x_i)	$u_i = \frac{(x_i - 136.95)}{10}$	$u_i\ f_i$
92 – 101.0	3	96.95	-4	-12
102 – 111.9	8	106.95	-3	-24
112 – 121.9	21	116.95	-2	-42
122 – 131.9	14	126.95	-1	-14
132 – 141.9	8	136.95	0	0
142 – 151.9	7	146.95	1	7
152 – 161.9	5	156.95	2	10
162 – 171.9	2	166.95	3	6
172 – 181.9	1	176.95	4	4
182 – 191.9	1	182.95	5	5
Total	70			-68

$$\mu_u = \frac{-68}{70} = -0.9714$$

$$\mu_x = -0.9714 \times 10 + 136.95 = 127.236$$

Note that it is convenient to take the class value of a class interval in the middle of the table as the base A, as we have done above.

The Arithmetic Mean is always used as a measure of central tendency, unless there is a specific reason for not doing so. One such reason is that it is too much affected by extreme values. So if the data show either some extremely high values or some extremely low values, the arithmetic mean may lose its representative character. Thus, consider the data 1,2,3,4,100. The arithmetic mean is 22, which is above four of the five values, and hence it will be misleading to take it as a representative value for the aggregate. Data from the economic field (for example, data on per capita income of families) may show this type of variation. In such cases we prefer to use some other measure of central tendency.

11.4.2 The Median

The median is defined as the middle value of a random variable. Thus, there are exactly as many values above the median as there are below. We now give the procedures for calculating the median from ungrouped and grouped data.

Calculation of Median from Ungrouped Data

We first arrange the data in ascending or descending order. If there is an odd number of values of the random variable, say N = 2M + 1, then the (M + 1)th value gives the median. If the number of values is even, say N = 2M, the Mth and (M + 1)th values in order of magnitude stand jointly in the middle and so their arithmetic mean is taken as the median.

Example 6 : The hours of relief obtained by 9 patients using an analgesic are given below.

3.2, 1.6, 5.7, 2.8, 5.5, 1.2, 6.1, 2.9, 5.8

Find the median.

Solution : Arranging the values in order of magnitude, we get

1.2, 1.6, 2.8, 2.9, 3.2, 5.5, 5.7, 5.8, 6.1

The middle value is the 5th value.

The 5th value in this data, 3.2 is the median.

Example 7 : The body temperature in °C of 8 patients in a hospital are 38.1, 37.4, 37.9, 37.9, 37.6, 37.8, 37.7, 38.1. Find the median.

Solution : Arranging the values in an ascending order we get

37.4, 37.6, 37.7, 37.8, 37.9, 37.9, 38.1, 38.1

The median is given by the mean of the 4th and 5th values, which lie in the middle. Therefore, the median $= \frac{37.8 + 37.9}{2} = 37.85$.

Now let us discuss the procedure for grouped data.

Calculation of Median from Grouped Data

To calculate the median from grouped data, we have to first draw up a table of cumulative frequencies of the random variable. To find the cumulative frequency upto a value in the data, we count the number of all the values in the data which are less than or equal to the given value. We'll illustrate this by a simple example. Consider the data

4, 2, 3, 1, 7, 1, 4, 4, 3, 8

Here the frequency of the value 4 is 3. And the cumulative trequency of 4 is 8, because there are 8 values in the data, namely, 4, 2, 3, 1, 1, 4, 4, 3 which are less than or equal to 4.

Table 9 shows the cumulative frequencies obtained from the frequency distribution of Table 5. The values of the random variable in the first column correspond to class boundaries.

Table 9

Systolic Blood Pressure (mm)	Frequency	Cumulative Frequency
91.95	0	0
101.95	3	3
111.95	8	11
121.95	21	32
131.95	14	46
141.95	8	54
151.95	7	61
161.95	5	66
171.95	2	68
181.95	1	69
191.95	1	70

To find the cumulative frequency of a particular class boundary, we first locate the class which has the given value as its upper boundary. Then we add the frequencies of all the classes which appear before that class in the table to the frequency of that class. This is the required cumulative frequency.

Now half of 70 is 35. So the thirtyfifth value in order of magnitude is the median. It is thus between 121.95 and 131.95. But how do we find it exactly? We now explain the procedure. Suppose the class boundaries of the different classes in the frequency distribution are denoted by x_1, x_2, x_3, ..., and the cumulative frequencies upto these values are F_1, F_2, F_3, ... Suppose $F_i < \frac{N}{2} < F_{i+1}$

for some i, where N is the total number of observations in the data. Then the median lies between x_i and x_{i+1}. If the length of the class interval is C and the frequency is f_i, then

$$\text{Median} = x_i + \frac{\left(\frac{N}{2} - F_i\right)}{f_i} \times C$$

Now applying this formula, we can calculate the median for the data in Table 9. For this case we have $N = 70$, $C = 10$, $x_i = 121.95$, $F_i = 32$ and $f_i = 14$.

$$\text{Thus, median} = 121.95 + \frac{35-32}{14} \times 10 = 124.093.$$

The value of the median is unaffected by extreme values of a random variable. Thus, it remains representative of an aggregate even if extremely high or low values are obtained by a few individuals. In this respect is has a distinct advantage over the mean, which makes it suitable for many types of economic data, particularly for computation of average income per capita.

In the next sub-section we'll discuss the third measure of central tendency – the mode.

11.4.3 Mode

The third commonly used measure of central tendency is the mode. It is the value which corresponds to the highest frequency of the variable. For a discrete variable, it can be computed exactly, either from ungrouped population data or from a frequency distribution of the population with single values as classes. In all other cases the computation of even its approximate value is impossible or difficult. For a discrete or a continuous random variable grouped in a frequency distribution with classes having more than one value, it is approximately given by

$$\text{Mode} = x_0 + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times C$$

where x_0 is the lower class boundary of the class interval with the highest frequency f_0; f_{-1} and f_1 are the frequencies of the immediately preceding and immediately following class intervals, and C is the length of the class interval. However, we would like to tell you that the formula is not at all satisfactory and depends too much on the choice of the class intervals.

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

(Mean–Median)is about $\frac{1}{3}$ (Mean – Mode). So

Mode = Mean – 3 (Mean – Median).

See if you can do these exercises now.

E7) Find the mean, median and mode from the following distribution, using both the above formulas for mode.

Weight (in mg) of dry contents of a certain type of ampoule

Weights	Class values	Frequency	Cumulative Frequency	Class value × Frequency
80 – 84		2		
85 – 89		7		
90 – 94		18		
95 – 99		32		
100 – 104		14		
105 – 109		10		
110 – 114		3		
115 – 119		1		
120 – 124		1		

E8) Suppose the mean weight of 100 males is 78 kgs. and that of 150 females is 72 kgs. What is the mean weight of all the 250 persons, males and females, taken together? (Hint : $\mu_m = 78$ and $\mu_f = 72$. If X denotes the combined variable, then

$$\sum_{i=1}^{250} x_i = \sum_{i=1}^{100} m_i + \sum_{i=1}^{150} f_i,$$

where M and F are the random variables representing the weights of males, and females, respectively.

11.5 MEASURES OF DISPERSION

At the end of Sec. 11.3 we talked about two important characters of a random variable : average and variability. We have discussed the average in the last section. Here we shall introduce the measures of the variability or dispersion of a random variable. These measures indicate the extent to which the individual values in the data are scattered around a typical value.

Among the measures of dispersion, the simplest one is the **range** = maximum value – minimum value. But range is not regarded as a good measure, since it is too unstable and does not take into account the scatter of the values in between the maximum and the minimum.

For example the range of 3, 4, 5, 6, 20 is 17, but in the absence of 20, the range becomes only 3.

The range of 1, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10, 21 is the same as that of 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, but you can see from Fig. 3 that the two variables have entirely different variability characteristics.

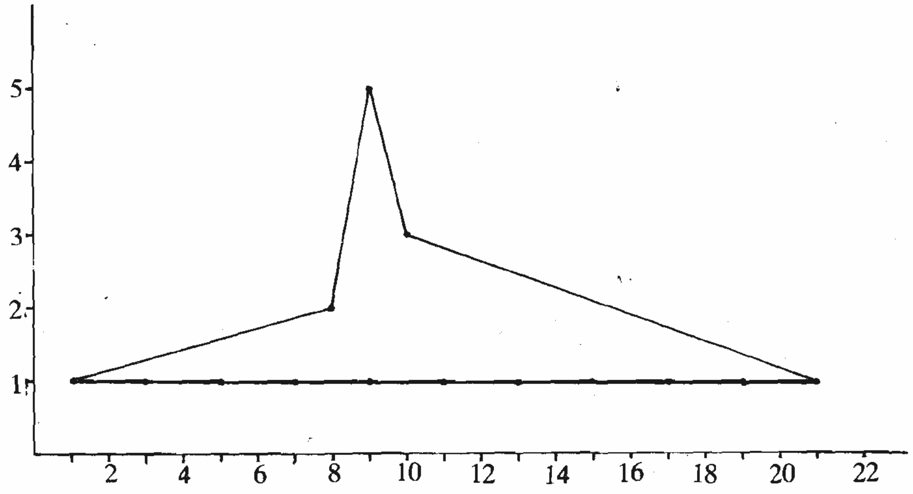


Fig. 3

Now we shall discuss three important measures of dispersion

(i) standard deviation, (ii) mean deviation, and (iii) semi-interquartile range. Let's tackle these one by one.

11.5.1 Standard Deviation

As before, we shall give the method of calculation for both ungrouped and grouped data.

Ungrouped Data : The standard deviation is obtained by squaring the deviations of the individual values from the mean, averaging these squares, and then extracting the square root to get back to the scale of the variable.

In other words, we take the following steps :

Step 1) Calculate the mean.

Step 2) Find the deviations of all the values from the mean.

Step 3) Square these deviations.

Step 4) Sum these squares.

Step 5) Divide by the total number of values.

Step 6) Take the square root.

We will now illustrate this procedure through an example.

Example 8 : Calculate the standard deviation of

3,6,8,12,8,6,15,8,9,7.

Solution : Step 1) Mean $\mu = \frac{82}{10} = 8.2.$

Step 2) The deviations of the values from the mean of the values are

$-5.2, -2.2, -.2, 3.8, -.2, -2.2, 6.8, -.2, .8, -1.2$

Step 3) The squares of the deviations are 27.04, 4.84, .04, 14.44, .04, 4.84, 46.24, .04, .64, 1.44.

Step 4) + 5) The average of these squares is 9.96

Step 6) Hence the standard deviation is $\sqrt{9.96} = 3.16$.

For ungrouped data $x_1, x_2, \ldots\ldots x_N$, the standard deviation is

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu^2)}$$

On simplification, σ reduces to

$$\sigma = \sqrt{\frac{1}{N} \sum x_i^2 - \mu^2}$$

Also, if $u_i = \frac{x_i - A}{C}$, then the standard deviations σ_x and σ_u of the random variables X and U are connected by the formula

$$\sigma_x^2 = C^2 \ \sigma_u^2 \qquad \ldots(I)$$

Grouped Data

For data grouped in a frequency distribution with class values $x_1, \ldots\ldots, x_k$ and frequencies $f_1, \ldots\ldots, f_k$, the standard deviation is given by

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{k} f_i \, (x_i - \mu)^2}, \quad \text{where N is the total frequency.}$$

This reduces to

$$\sigma = \sqrt{\frac{\sum f_i \, x_i^2}{N} - \mu^2}$$

If u_i's are obtained from x_i's by making the same transformation of base and scale as before, the standard deviations of X and U are again connected by formula (I). As in the case of mean, we follow the "short method" to calculate the standard deviation from a frequency distribution if the class intervals are equal.

Example 9: The first two columns of Table 10 give the frequency distribution of weights (in gms.) of sample packets of seed.

Compute the standard deviation of X = Weight (in gm.) of seed samples by taking A = 6.245, C = 0.5

Table 10: Weight (in gms.) of sample packets of seed

Class values (x_i)	Frequency (f_i)	$u_i = \frac{x_i - 6.245}{0.5}$	$u_i \, f_i$	$u_i^2 \, f_i$
3.245	1	-7	-7	49
3.745	1	-6	-6	36
4.245	1	-5	-5	25
4.745	3	-4	-12	48
5.245	7	-3	-21	63
5.745	8	-2	-16	32
6.245	18	-1	-18	18
6.745	13	0	0	0
7.245	7	1	7	7
7.745	4	2	8	16
8.245	0	3	0	0
8.745	3	4	12	48
Total	66		-58	342

Solution: Here

$$\mu_u = \frac{1}{N} \sum f_i \, u_i = \frac{-58}{66} = -0.8788;$$

Therefore $\mu_x = (0.5)\,(-0.8788) + 6.745 = 6.306$

$$\sigma_u^2 = \frac{1}{N} \sum f_i \, u_i^2 - \mu_u^2 = \frac{342}{66} - (.8788)^2 = 5.1818 - 0.7723 = 4.4095$$

$\sigma_u = 2.0999$, $\sigma_x = .5 \times 2.0999 = 1.05.$

Proceeding similarly you will be able to solve this exercise.

E 9) Calculate the standard deviation of head-lengths given in Table 8.

The absolute value of a number x is

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$$

11.5.2 Mean Deviation

Another measure of dispersion is obtained by averaging the absolute deviations of the values of a random variable X from some measure of central tendency.

Thus we can have mean deviation about the mean μ, or about the median or mode.

Ungrouped Data

Suppose we want to compute the mean deviation about the mean. We will have to take the following steps:

1) Calculate the mean
2) Find the deviations of all the values from the mean
3) Take the absolute value of these deviations
4) Add the absolute values
5) Divide by the total number of values

If you want to calculate the mean deviation about the median (or the mode), all you have to do is replace mean by median (or mode) in steps 1) and 2). The rest of the procedure remains the same. You can check that the mean deviation about the mean of the data given in Example 9 of the previous section is

$(5.2 + 2.2 + .2 + .. + 1.2)/10 = 2.28$. The median of the value is 8.

So the mean deviation about the median is

$$\frac{1}{10}(5 + 2 + 0 + 4 + 0 + 2 + 7 + 0 + 1 + 1) = 2.2$$

The mean deviation of the values $x_1, \ldots\ldots., x_N$, of X about a measure of central tendency, A, is given by

$$MD_A = \frac{1}{N}\sum_{i=1}^{N} | x_i - A |.$$

Grouped Data: The formula for mean deviation about a measure of central tendency, A, for a grouped data is

$MD_A = \frac{1}{N}\sum f_i \, | x_i - A |$, where x_i are the class values, f_i are the corresponding frequencies, and N is the total frequency.

Example 10: Find the mean deviation about the mean of the frequency distribution given in Table 10.

Solution: We proceed as follows. $\mu_x = 6.306$

Class value x_i	Frequency f_i	$x_i - \mu_x$	$f_i \mid x_i - \mu_x \mid$
3.245	1	3.061	3.061
3.745	1	2.561	2.561
4.245	1	2.061	2.061
4.745	3	1.561	4.683
5.245	7	1.061	7.427
5.745	8	.561	4.488
6.245	18	.061	1.098
6.745	13	.439	5.707
7.245	7	.939	6.573
7.745	4	1.439	5.756
8.245	0	1.939	0
8.745	3	2.439	7.317
Total	66		50.732

$$\text{MD about mean} = \frac{50.732}{66} = 0.769.$$

Now we turn our attention to the third measure of dispersion.

11.5.3 Semi-interquartile Range

Another measure of dispersion is given by the **semi-interquartile range,** also known as the **quartile deviation.** A value of X below which 25% of the values lie is called the **first quartile** and is denoted by Q_1.

A value of X below which 75% of the values of X lie is called the **third quartile,** and is denoted by Q_3. The median is also called the second quartile because 50% of the values lie below it. The semi-interquartile range is given by

$$Q = \frac{Q_3 - Q_1}{2}.$$

The values Q_1 and Q_3 can be calculated in a manner similar to the median for grouped or ungrouped data.

For example, if N is the total number of observations, then there are $\frac{N}{4}$ values of X below Q_1. This means we can locate the class in which Q_1 lies with the help of cumulative frequencies. Suppose Q_1 lies in the class whose lower boundary is x_i, and whose frequency is f_i. Suppose further that F_i is the cumulative frequency of x_i,

Then,

$$Q_1 = x_i + \frac{(N/4) - F_i}{f_i} \times C,$$ where C is the length of the class interval.

Similarly you should be able to argue that

$$Q_3 = x_j + \frac{(3N/4) - F_j}{f_j} \times C$$

where x_j, f_j, F_j and C have their relevant meanings.

The procedure will become clear if you read the next example.

Example 11: From the frequency distribution of head lengths given in Table 8 obtain the cumulative frequency distribution and hence Q.

Solution: The cumulative frequency distribution is given by

Class boundaries	Cumulative frequency
171.5	0
175.5	3
179.5	12
183.5	41
187.5	117
191.5	221
195.5	331
199.5	419
203.5	449
207.5	455
211.5	459
215.5	461
219.5	462

$$\frac{N}{4} = \frac{462}{4} = 115.5;$$

$$\therefore Q_1 = \frac{115.5 - 41}{76} \times 4 + 183.5 = 187.5$$

$$\frac{3N}{4} = \frac{3 \times 462}{4} = 346.5$$

$$\therefore Q_3 = \frac{346.5 - 331}{88} \times 4 + 195.5 = 196.2$$

$$\text{Thus, } Q = \frac{Q_3 - Q_1}{2} = 4.35$$

Try to do this exercise now.

E10) Compute Q_1 and the M.D. about the median for the following distribution.

Grades of 240 participants in a Quiz contest

Grade	Frequency
0 – 9	1
10 – 19	3
20 – 29	7
30 – 39	11
40 – 49	17
50 – 59	23
60 – 69	32
70 – 79	60
80 – 89	51
90 – 99	35

Now before ending this unit, we'll summarise what we have covered in it.

11.6 SUMMARY

In this unit we have

1) introduced terms like statistical data, discrete and continuous random variables, individuals, population, sample,
2) discussed frequency distributions for discrete and continuous random variables,
3) seen how to represent the given data by frequency polygons and histograms,
4) given the methods of calculation for three important measures of central tendency, namely, the mean, median and mode, for ungrouped and grouped data,
5) discussed the computation of measures of dispersion like the standard deviation, the mean deviation and the semi-interquartile range for ungrouped and grouped data.

11.7 SOLUTIONS AND ANSWERS

E1) b) and d) are discrete; a), c) and e) are continuous.

E2) a) A family residing in the locality is the individual, and the size of the family is the variable.

b) percentage of alcohol is the variable, and a wine bottle of any of the 20 brands is the individual.

c) Volume of sales of TV sets in a year is the variable, and a year is the individual.

E3)

No. of peas	Tally marks	Frequency
1	||	2
2	𝍸 𝍸 𝍸 𝍸 𝍸 ||	27
3	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	53
4	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	45
5	𝍸 𝍸 𝍸 |	16
6	𝍸 |	6
7	|	1
Total		150

E4) a) X is discrete. The minimum score is 34, and the maximum is 96.

b)

Class interval	Tally marks	Frequency
30 – 39	||	2
40 – 49	|||	3
50 – 59	𝍸 𝍸 |	11
60 – 69	𝍸 𝍸 𝍸 𝍸	20
70 – 79	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||	32
80 – 89	𝍸 𝍸 𝍸 𝍸 𝍸	25
90 – 99	𝍸 ||	7

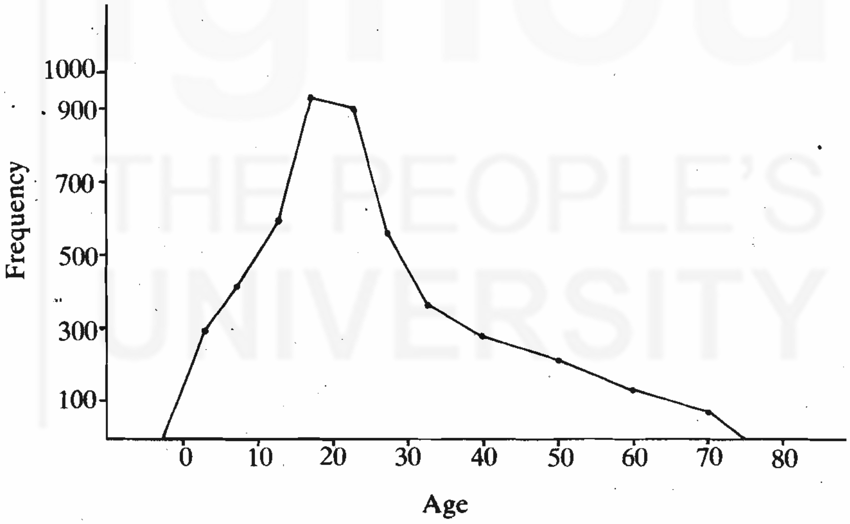
c) 0.95

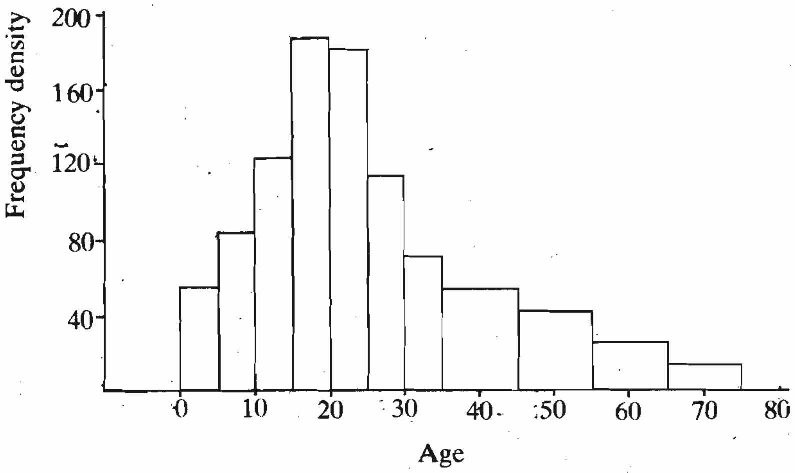
d) 0.32

e) 0.63

f) The number of students scoring between 55 and 59 is approximately equal to $\frac{11}{2}$. The number of students scoring between 70 and 74 is about 16. The required number is 41.5.

E5)





E6) $$\sum_{i=1}^{N} (x_i - \mu) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu$$

$$= \sum_{i=1}^{N} x_i - N\mu$$

$$= \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0.$$

E7)

Weights	Class	Frequency	Cumulative Frequency	Class value × Frequency
80 – 84	82	2	2	164
85 – 89	87	7	9	609
90 – 94	92	18	27	1656
95 – 99	97	32	59	3104
100 – 104	102	14	73	1428
105 – 109	107	10	83	1070
110 – 114	112	3	86	336
115 – 119	117	1	87	117
120 – 124	122	1	88	122
Total		88		8606

Mean $= \frac{8606}{88} = 97.795$

Median is the $\frac{88}{2}$ = 44th value in order of magnitude. Therefore, it is in the class interval 94.5 - 99.5.

$\therefore$ Median $= 94.5 + \frac{44-27}{32} \times 5 = 97.15$

Model class, i.e., the class with the highest frequency is 94.5 - 99.5

$$\text{Mode} = 94.5 + \frac{(32-18)}{64-18-14} \times 5$$

$$= 94.5 + 2.1875$$

$$= 96.69$$

Or, mode = 97.795 - 3 (97.795-97.15) = 95.86

E8) Total weight of 100 males with mean weight 78 kg. is 78 × 100 = 7800 kg.

Total weight of 150 females with mean weight of 72 kg. is 72 × 150 = 10800 kg.

Total weight of 250 persons (100 males + 150 females) is 7800 + 10800 = 18600 kg.

Mean weight $= \frac{18600}{250} = 74.4$ kg.

E9)

Class value x_i	Frequency f_i	$u_i = x_i - 195.5$	$f_i\ u_i$	$f_i.\ u_i^2$
173.5	3	-22	-66	1452
177.5	9	-18	-162	2916
181.5	29	-14	-406	5684
185.5	76	-10	-760	7600
189.5	104	-6	-624	3744
193.5	110	-2	-220	440
197.5	88	2	176	352
201.5	30	6	180	1080
205.5	6	10	60	600
209.5	4	14	56	784
213.5	2	18	36	648
217.5	1	22	22	484
Total	460		-1726	25784

$$\mu_u = \frac{-1726}{460} = -3.7$$

$$\sigma_u^2 = \frac{1}{N} \Sigma f_i \; u_i^2 - \mu_u^2$$

$$= \frac{25784}{460} - (3.7)^2$$

$$= 56.05 - 13.69 = 42.36$$

$\therefore \sigma_u = 6.508 \qquad \therefore \sigma_x = 6.508$ since here C = 1.

E10)

Class Interval	Frequency f_i	Cumulative Frequency	Class Value x_i	$\mid x_i - Q_2 \mid$	$f_i \mid x_i - Q_2 \mid$
0.5–9.5	1	1	5	56.375	56.375
9.5–19.5	3	4	15	46.375	139.125
19.5–29.5	7	11	25	36.375	254.625
29.5–39.5	11	22	35	26.375	290.125
39.5–49.5	17	39	45	16.375	278.375
49.5–59.5	23	62	55	6.375	146.625
59.5–69.5	32	94	65	3.625	116.000
69.5–79.5	60	154	75	13.625	817.500
79.5–89.5	51	205	85	23.625	1204.875
89.5–99.5	35	240	95	33.625	1176.875
	240				4480.500

$$\frac{N}{4} = 60, \; \frac{N}{2} = 120$$

Q_1 and the median correspond to the 60th value and the 120th value in order of magnitude, respectively.

Q_1 lies in the interval 49.5-59.5, and the median lies in the interval 69.5-79.5.

$$Q_1 = 49.5 + \frac{60-39}{23} \times 10 = 50.413$$

$$Q_2 = \text{Median} = 69.5 + \frac{120-94}{32} \times 10 = 61.375$$

$$\text{M.D. about } Q_2 = \frac{4480.5}{240} = 18.67$$