
UNIT 1 DESCRIPTIVE STATISTICS

Structure	Page No.
1.1 Introduction Objectives	7
1.2 Collecting Data Kinds of Data Frequency Distribution of a Variable Graphical Representation of Frequency Distributions	8
1.3 Summarisation of Data Measures of Central Tendency Measures of Dispersion or Variability	21
1.4 Summary	32
1.5 Solutions/Answers	33

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?

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.

Population is a collection of all the individuals we are studying

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,

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% 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 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.
- All the adult males residing in India.
 - Twenty cancer patients chosen to participate in a program to test a new drug.
 - 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?
- age
 - year of birth
 - height
 - number of students admitted to the school in a calendar year

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	
4	
5	
6	
7	
8	
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	Cumulative Frequency
	(Less than type)	(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	
750.6-830.5	//
830.6-910.5	//
910.6-990.5	/
990.6-1070.5	
1070.6-1150.5	//
1150.6-1230.5	/
1230.6-1310.5	/
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
- the class limits
 - the class boundaries
 - the class marks
 - 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.

- Is the random variable X = Score of a student, discrete or continuous? What are the minimum and maximum scores?
- Using the classes 20 – 29, 30 – 39, 40 – 49, ... and 90 – 99 draw up the frequency distribution of X .
- What percentage of the students score above the pass marks of 50?
- 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.

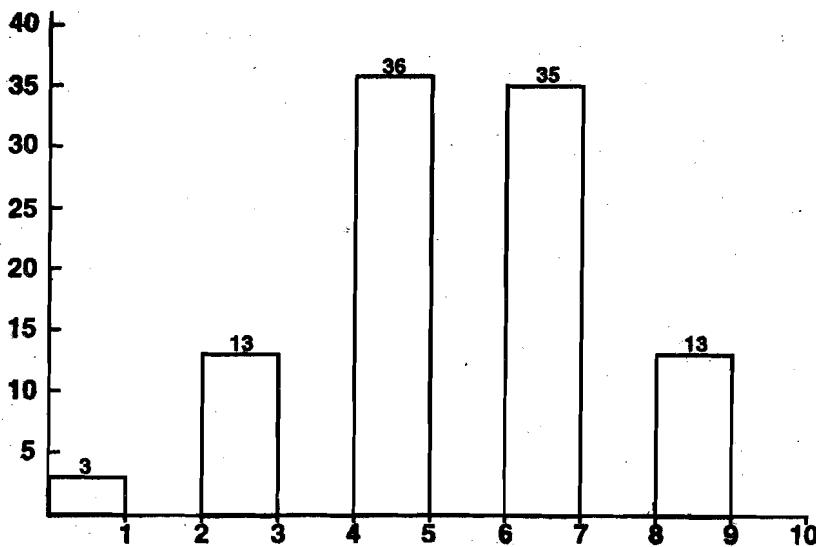


Fig.1

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.

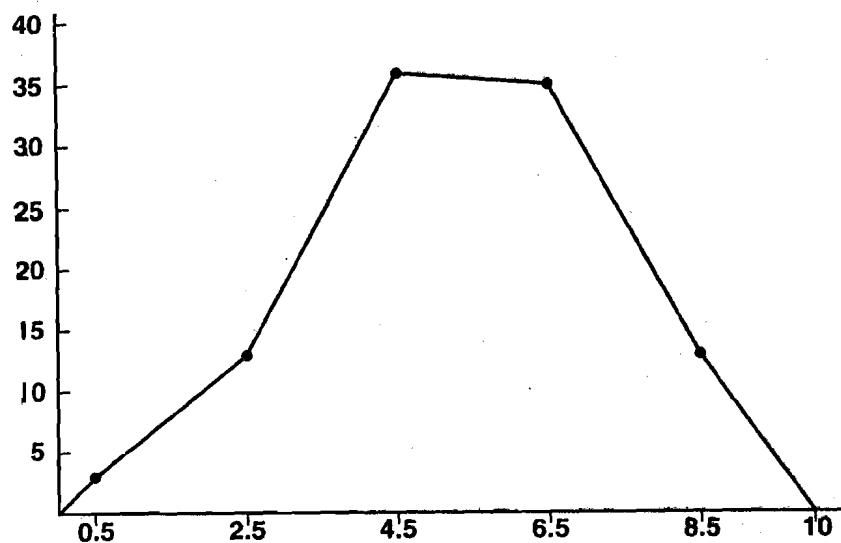


Fig.2

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.

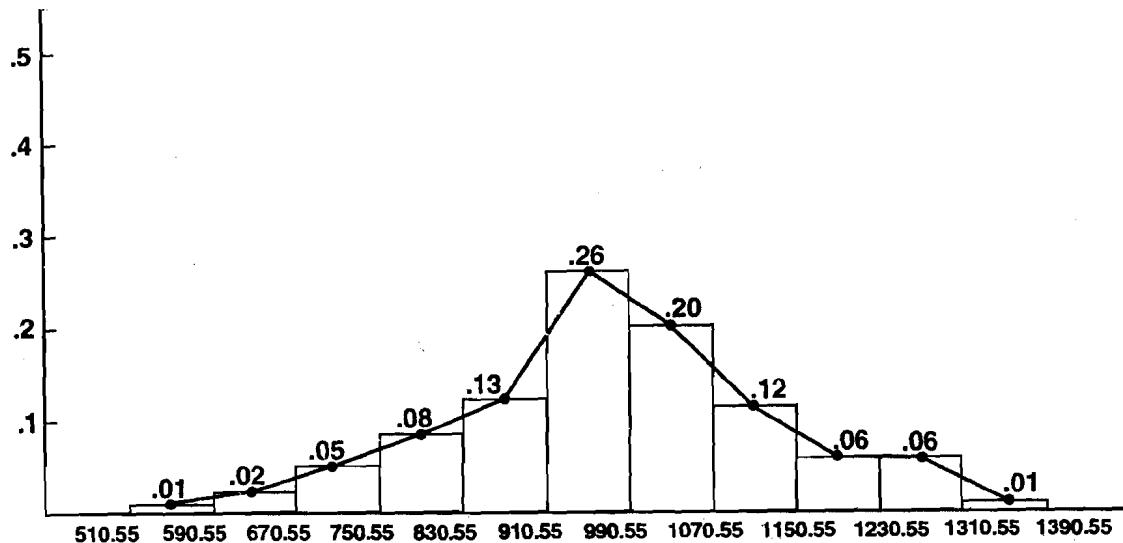


Fig.3

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.

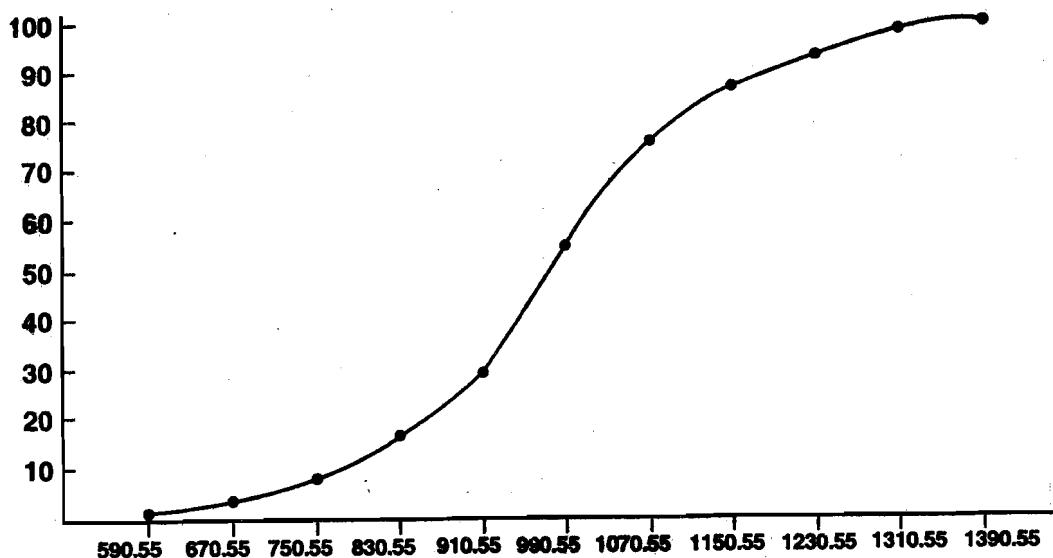


Fig.4

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.

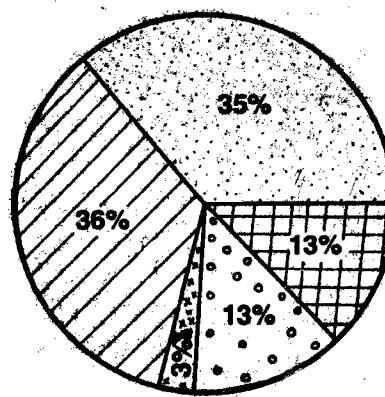


Fig.5

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 N th number, then the population mean is

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

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, \dots, x_k with f_1, f_2, \dots, f_k as the corresponding frequencies (and thus k is the number of classes). Since each x_i appears f_i times, $i = 1, 2, \dots, k$ in the data, the sum of all observations is equal to

$$\begin{aligned} & x_1 + x_1 + \dots + x_1 \quad + x_2 + x_2 + \dots + x_2 \quad + x_k + x_k + \dots + x_k \\ & \leftarrow f_1 \text{ times} \rightarrow \quad \leftarrow f_2 \text{ times} \rightarrow \quad \leftarrow f_k \text{ times} \rightarrow \\ & = x_1 f_1 + x_2 f_2 + \dots + x_k f_k \end{aligned}$$

Now the arithmetic mean will be given by

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

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

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 i th measurement in a sample, the **weighted arithmetic mean** is

$$\bar{x}_w = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i} \quad (3)$$

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, 7 and 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

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

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

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

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

$$\begin{aligned} (\text{Mean} - \text{Median}) &= \frac{1}{3}(\text{Mean} - \text{Mode}) \\ \therefore \text{Mode} &= \text{Mean} - 3(\text{Mean} - \text{Median}) \end{aligned}$$

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

- The value of every observation in the data set is taken into account when we calculate its median.
- Measure of central tendency in a data set refer to the extent to which the observations are scattered.
- We can compute a mean for any data set, once we are given its frequency

- 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: 1,1,1,1,1,1,1,1,10

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, \dots, x_N be N observations and let μ be the mean of these observations. Consider the deviations $(x_j - \mu), j = 1, 2, \dots, N$ of the observations from the mean. The variation or dispersion is small if the observations x_1, x_2, \dots, 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.

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

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

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

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

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

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

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, \dots, x_k are the distinct observations, f_j is the frequency of x_j and $f_1 + f_2 + \dots + f_k = N$, 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, $j = 1, 2, \dots, k$. 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:

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

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

$$\begin{aligned} \sigma^2 &= (1/100) \{ 3219 - (100)(5.35)^2 \} \\ &= 3.5675 \end{aligned}$$

$$\text{so that } SD = (3.5675)^{\frac{1}{2}} = 1.8888.$$

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 $x_j, j = 1, 2, \dots, k$ and f_j be the frequency of x_j . Let us write

$y_j = (x_j - a), j = 1, 2, \dots, k$, where a and b are two pre-specified constants.

Thus, $x_j = a + b y_j, j = 1, 2, \dots, k$.

Hence,

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

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

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

so that, we obtain the following result.

$$\text{Variance of x-observations} = b^2 (\text{Variance of y-observations}), \quad (11)$$

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

$$\begin{aligned} \mu' &= \sum_{i=1}^{11} \frac{y_i f_i}{N} \\ &= \frac{34}{100} = 0.34 \end{aligned}$$

and the estimated mean μ of the x-observations is

$$\begin{aligned} \mu &= a + b \mu' \\ &= 950.55 + 80(0.34) \\ &= 977.75 \text{ hours.} \end{aligned}$$

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

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^{11} y_j^2 f_j - \mu'^2 &= \frac{351}{100} - (0.34)^2 \\ &= 3.39 \end{aligned}$$

Now using Formula (11),

$$\begin{aligned} \text{the estimated variance of } x\text{-observations} &= \sigma^2 = b^2(\text{variance of } y\text{-observations}) \\ &= (80)^2(3.39) = 21696(\text{hours})^2 \\ \therefore \text{the estimated SD of the } x\text{-observations} &= \sigma = +\sqrt{21696} = 147.2956 \text{ hours.} \end{aligned}$$

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, \dots, 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, \dots, y_m are \bar{y} and σ_y respectively. Show that the standard deviation of the pooled set $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m$ of $m+n$ observations is

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

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:

$$\begin{aligned} & \frac{100}{1000} \{ \text{Total frequency} - \text{frequency in } (0, 4000) \\ & \quad - \frac{152}{6000 - 4000} (4500 - 4000) \} \\ & = \frac{100}{1000} \{ 1000 - 236 - 38 \} = 72.6\% \end{aligned}$$

ii) As above, an estimate is

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

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

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.

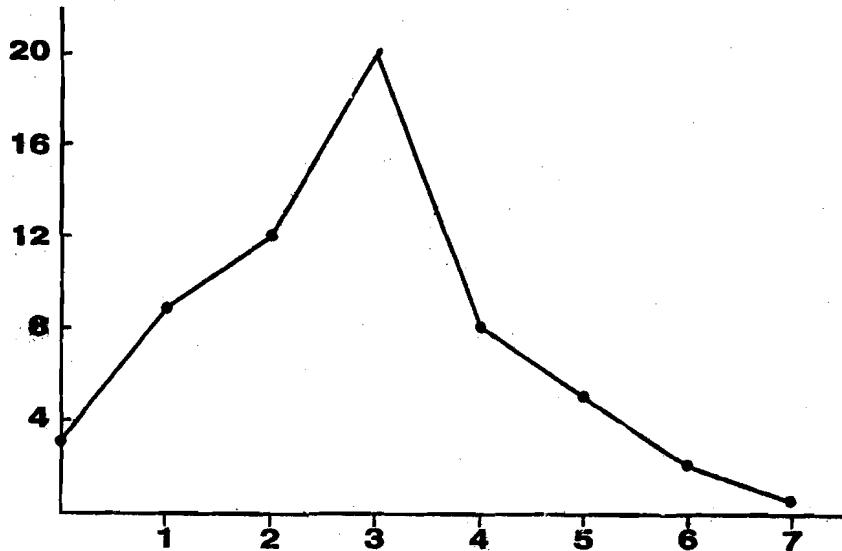


Fig.6

Similarly draw cumulative frequency graph for the above data.

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

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

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 = $\{60(1.15)(60) + 40(.80)(75)\} / 100 = 65.40$

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

$$\text{E14) For the given data } \frac{n}{2} = \frac{70}{2} = 35$$

$c = 31, f_m = 16, l = 24.9, L_m = 75$, using Formula (4), median $M = 81.23$

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

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

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

- 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	$y^2 f$
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

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

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

UNIT 2 PROBABILITY CONCEPTS

Structure	Page No.
2.1 Introduction Objectives	36
2.2 Preliminaries Trials, Sample Space, Events Algebra of Events	37
2.3 Probability Concepts Probability of an Event Probability of Compound Events	42
2.4 Conditional Probability and Independent Events	50
2.5 Summary	54
2.6 Solutions/Answers	54

2.1 INTRODUCTION

Sunil is an enterprising class VIII student who wants to use his free hours more fruitfully. A news paper agent agrees to employ him for one hour in the morning between 5.30 AM to 6.30 AM for distributing news papers in a residential colony where there are 85 regular subscribers. In addition, Sunil finds that there are about 10 irregular customers who may buy the paper from him on a day to day basis. On every additional news paper Sunil sells, he makes an extra income of 30 paise. But on every unsold news paper that he takes back to the agent he loses 10 paise. Sunil has to decide how many newspapers he should collect from the agent each morning so that he makes the maximum possible gain. His dilemma is about the 10 irregular customers as he has to decide how many of these 10 will actually buy from him on any given day. If you ask him whether he knows probability theory, he may be surprised and say he knows nothing. Actually he may already be using some of the ideas of probability theory while not being aware of or articulate about them. This is a very commonly occurring situation.

When asked, "Do you know anything about probability?" most people are quick to answer, "No!" Usually that is not the case at all. The words probable and probably are used commonly in everyday language. We say, "It will probably rain tomorrow" or "there is 0% chance of rain today." Such statements often have a vague, subjective quality and based sometimes on certain information and at other times on intuition only.

We live in an uncertain world. When we get up in the morning we cannot say exactly who we are going to meet, what the weather will be like or what event will be on the television news during the day.

In our everyday lives, we cope with this uncertainty by making hundreds of guesses, calculated risks and some gambles. We don't take a coat with us for the weekend because it is unlikely to be cold. We allow a particular length of time to travel to an important interview because it will probably be enough.

All these decisions are made by assessing the relative probability (chance) of all the possible outcomes -- even if we do this unconsciously and intuitively. Business decisions are made in a similar climate of uncertainty. A publisher must decide how large the print run of a new book should be to avoid unusual storage of unsold copies and yet ensure availability. A stock market dealer decides to sell a particular share because a financial model tells her that the price is likely to fall.

The penalties of estimating chances inaccurately and hence making a wrong decision vary from minor inconvenience, to loss of income to bankruptcy. So, in business (and other fields) we endeavour to measure uncertainty using some scientific method. Rather than make vague statements containing 'likely', 'may be' or 'probably'; we need to be more precise.

Historically, the oldest way of measuring uncertainties is the probability concept. Probability theory had its beginnings over 300 years ago, when gamblers of that period asked mathematicians to develop a system for predicting outcomes of a turn of the roulette wheel or a roll of a pair of dice. The word probability is associated with a quantitative approach to predicting the outcome of an event (the outcome of a presidential election, the side effects of a new medication, etc.).

In this unit we shall see how uncertainties can actually be measured, how they can be assigned numbers (called probabilities) and how these numbers are to be interpreted. After starting with some preliminaries of the probabilities we shall concentrate on the rules which probabilities must obey. This includes the basic postulates, the relationship between probabilities and odds, the addition rules, the definition of conditional probability, the multiplication rules etc.

Objectives

After reading this unit, you should be able to

- describe trials, events, sample spaces associated with an experiment;
- express the union, intersection, complement of two or more events in terms of a new event;
-] define the probability of the occurrence of an event and obtain it;
- obtain the conditional probability of an event.

2.2 PRELIMINARIES

What is an experiment? Many of you will relate experiments with all that you were expected to do in physics, chemistry or biology laboratories in your schools and colleges. For example, you may perhaps recollect that in the chemistry laboratory, one of the experiments you performed was to explore as to what would happen if sulfuric acid is poured in a jar containing zinc. Yet in another experiment in the physics laboratory, you might have performed the act of inserting a battery of a certain specification in a given circuit with a view to finding out the quantum of the flow of electricity through the said circuit. Thus, generally speaking, an **experiment** is merely the performance of an act for generating an observation on the phenomenon under study. In fact, when you pick up an item from a lot consisting of a number of items coming out of a manufacturing process, say, to decide whether the picked up item is defective or not, then also you are performing an experiment. Whenever you are investing a certain sum of money in the share market to see to what extent your money grows over a certain specified time interval, you are performing an experiment too.

When you are stocking a number of units of a particular brand of a consumer good in your store in anticipation of sale, that is also an experiment. Tossing a coin, rolling a die, observing the number of road accidents on a given day in a city etc. are all experiments. In each such case, a certain act is performed and its outcome is observed. Is there then any difference between the former type of laboratory experiments and the latter type of experiments that we presumably perform in some form or the other in our daily lives? The answer to this question is 'Yes'. The difference is in terms of the outcomes that you associate with the experiments. Notice that in the former type of experiments performed under the controlled conditions of a laboratory, the outcomes

are known a priori from the conditions under which these are carried out. It is known apriori that sulfuric acid and zinc together will yield hydrogen gas and zinc sulfate. The quantum of electricity flow through the circuit is known from the specifications of the inserted battery through the celebrated Ohm's Law. Such experiments are **deterministic** in the sense that the conditions under which these are carried out would inform us what the result is going to be even before the experiment is performed. However, when you are blindly picking up an item from a manufactured lot, you have no way of knowing a priori whether the item being picked up would be defective or not. When in your role as a shopkeeper, you stock certain units of a particular commodity, you will have no prior knowledge what your sales will be like over a specified period. When you toss a coin (it is equivalent to picking up an item blindly from a lot containing both defective and non-defective items - how?), you do not know the result of the experiment beforehand. Experiments whose outcomes cannot be precisely predicted primarily because the totality of the factors influencing the outcomes are either not precisely identifiable or not controllable at the time of experimentation, even if known, are called **random** or **stochastic** experiments. If you look at the dilemma of Sunil, the news paper boy, do you think what Sunil observes regarding the buying behaviour of the ten irregular customers on a given day can be modelled as the outcome of a random experiment? As Sunil does not know whether any of the specified irregular customer will actually buy from him, it is possible to think of this as a random experiment.

For the purposes of this block, by an experiment, we shall always mean a random experiment only. We now introduce some of the commonly occurring terms of the probability theory.

2.2.1 Trials, Sample Space, Events

You must have often observed that a random experiment may comprise of a series of smaller sub-experiments. These are called **trials**. Consider for instance the following situations.

Example 1: Suppose the experiment consists of observing the results of three successive tosses of a coin. Each toss is a trial and the experiment consists of three trials so that it is completed only after the third toss (trial) is over.

* * *

Example 2: Suppose from a lot of manufactured items, ten items are chosen successively following a certain mechanism for checking. The underlying experiment is completed only after the selection of the tenth item is completed; the experiment obviously comprises of 10 **trials**.

* * *

Example 3: If you consider Example 1 once again you would notice that each toss (trial) results into either a head (H) or a tail (T). In all there are 8 possible outcomes of the experiment viz., $s_1 = (H,H,H)$, $s_2 = (H,H,T)$, $s_3 = (H,T,H)$, $s_4 = (T,H,H)$, $s_5 = (T,T,H)$, $s_6 = (T,H,T)$, $s_7 = (H,T,T)$ and $s_8 = (T,T,T)$.

Each of the above outcome s_1, s_2, \dots, s_8 is called a **sample point**. Notice that each sample point has three entries separated by commas. For example, the sample point s_2 indicates that the first toss results in a head, the second also produces head while the third one leads to a tail. Thus the sequence in which H's and T's appear is important. The set $\zeta = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\}$ of all possible outcomes is called the **sample space** of the experiment. Sample spaces are classified as **finite** or **infinite** according to the number of sample points (finite/infinite) they contain. For instance, an infinite sample space arises when we throw a dart at a target and there is a continuum of points we may hit. Sample space that arises in the case of Sunil, the news paper boy is an example of a finite sample space. What do you think is the sample space in this case? In this case the random experiment is to observe the 10 irregular customers on any day

and note down whether each one "buys" or "does not buy" the newspaper. Therefore the sample space contains 2^{10} simple points which are sequences of the from (B, NB, NB, B, B, B, NB ...) etc. The sample space is, of course, finite. A specific collection or subset of sample points, say $E_1 = \{s_2, s_3, s_4\}$ is called an **event**. Event is a **simple event** if it consists of a single sample point. In this case $\{s_i\}$ is a simple event for $i = 1, 2, \dots, 8$. A **word of caution** here that not every subset of a sample space is an event. We shall not explain it here as it is beyond the scope of this course.

* * *

The following examples will further strengthen your understanding of various terms defined.

Example 4: Suppose our experiment consists in observing the number of road accidents in a given city on a given day. Obviously, $\zeta = \{0, 1, 2, \dots, b\}$, where b is the maximum possible number of accidents in a day and this can very well be infinity, in which case the sample space is infinite. The event E that there are five or less number of accidents on that day can be described by $E = \{0, 1, 2, 3, 4, 5\}$.

* * *

Example 5: Suppose, we are interested in noting down the time (in hours) to failure of an equipment. Here, the sample space $\zeta = \{x | 0 < x \leq b\} =]0, b]$, the half open interval between 0 and b, where b is the maximum possible life of the equipment. Here also b may be infinite. In any case, ζ is not only infinite, but also uncountable i.e. the elements of ζ cannot be put into one-to-one correspondence with the set of natural numbers. The event E that the equipment survives for at least 500 hours of operation can be described as $E = \{x | 500 \leq x \leq b\} = [500, b]$.

* * *

We shall now consider an example which shows the importance of order of selection of items/objects under consideration.

Example 6: Suppose an urn contains three marbles which are identical in all respect excepting that two of them are red in colour while the third one is white. The experiment is to pick up two marbles, one after the other, blindly and observing their colour. This experiment comprises of two trials. How does the experimenter report the result of the experiment? For this purpose, we may identify the two red marbles as r_1, r_2 and the white marble may be identified as w. Thus, a typical outcome, for instance, can be (w, r_2) ; this means that the white marble was picked up at the first draw while the second red marble appeared in the next draw. Note that the order in which w and r_2 appeared in (w, r_2) is important; the first letter corresponds to the first selection while the second letter corresponds to the second. Thus the sample space ζ of this random experiment is the following set:

$$\zeta = \{(r_1, r_2), (r_1, w), (r_2, r_1), (r_2, w), (w, r_1), (w, r_2)\}$$

How do we describe the 'event E that' of the two marbles picked up, one is red and the other is white'? Well, this event materialises if one of the two red marbles is picked up in one of the two trials while the white marble is picked up in the other trial. Thus, E materialises if the outcome of the experiment is either (r_1, w) or (r_2, w) or (w, r_1) or (w, r_2) ; thus, $E = \{(r_1, w), (r_2, w), (w, r_1), (w, r_2)\}$.

Let us now modify the conditions of the experiment a little. Suppose, we return the first marble selected back to the urn before the second marble is selected. Then, there will be some new possibilities. Specifically, now it becomes possible for the same marble to be selected twice. Hence, the sample space gets modified to

$$\zeta = \{(r_1, r_1), (r_1, r_2), (r_1, w), (r_2, r_1), (r_2, r_2), (r_2, w), (w, r_1), (w, r_2), (w, w)\}$$

but, the above event E remains unaltered.

Incidentally, in our initial experiment, marbles were being drawn **without replacement** while in the modified experiment, the marbles were being drawn **with replacement**.

Probability and Statistics

Selection is said to be done **with/without replacement** if each object chosen is **returned/not returned** to the population before the next object is drawn

The former experiment is equivalent to drawing two marbles in succession without the previously drawn marble being returned to the urn before the next draw. In the latter experiment the previously drawn marble was being replaced before the next draw, so that the number of marbles in the urn was always the same and it is possible that same marble is chosen more than once.

* * *

You may now try the following exercises:

E1) In each of the following exercises, an experiment is described. Specify the relevant sample spaces:

- A machine manufactures a certain item. An item produced by the machine is tested to determine whether or not it is defective.
- An urn contains six balls, which are coloured differently. A ball is drawn from the urn and its colour is noted.
- A person is asked to which of the following categories he belongs: unemployed, self-employed, employed by the Central Government, employed by a state or local Government, employed by a private organisation, employed by a employer who does not fulfil any of the above descriptions. The person's answer is recorded.
- An urn contains ten cards numbered 1 through 10. A card is drawn, its number noted and the card is replaced. Another card is drawn and its number is noted.

E2) Suppose a six-faced die is thrown twice. Describe each of the following events:

- The maximum score is 6.
- The total score is 9.
- Each throw results in an even score.
- Each throw results in an even score larger than 2.
- The scores on the two throws differ by at least 2.

It must have been quite clear to you by now that events associated with a random experiment can always be represented by the collection of sample points each of which leads to the occurrence of the said event. The phrase 'event E occurs' is an alternative way of saying that the outcome of the experiment has been one of those that lead to the materialisation of said event. Let us now see how an event can be expressed in terms of two or more events by forming unions, intersections and complements.

2.2.2 Algebra of Events

Let ζ be a fixed sample space. We have already defined an event as a collection of sample points from ζ . Imagine that the (conceptual) experiment underlying ζ is being performed. The phrase "the event E occurs" would mean that the experiment results in an outcome that is included in the event E. Similarly, non-occurrence of the event E would mean that the experiment results into an outcome that is not an element of the event E. Thus, the collection of all sample points that are not included in the event E is also an event which is complementary to E and is denoted as E^c . The event E^c is therefore the event which contains all those sample points of ζ which are not in E. As such, it is easy to see that the event E occurs if and only if the event E^c does not take place. The events E and E^c are **complementary events** and taken together they comprise the entire sample space, i.e., $E \cup E^c = \zeta$.

You may recall that ζ is an event which consists of all the sample points. Hence, its complement is an empty set in the sense that it does not contain any sample point and is called the **null event**, usually denoted as ϕ so that $\zeta^c = \phi$.

Let us once again consider Example 3. Consider the event E that the three tosses produce at least one head. Thus, $E = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\}$ so that the complementary event $E^c = \{s_8\}$, which is the event of not scoring a head at all.

Again in Example 6 in the case of selection without replacement, event that the white marble is picked up at least once is defined as $E = \{(r_1, w), (r_2, w), (w, r_2), (w, r_1)\}$. Hence, $E^c = \{(r_1, r_2), (r_2, r_1)\}$ i.e. the event of not picking the white marble at all.

Let us now consider two events E and F. We write $E \cup F$, read as E “union” F, to denote the collection of sample points, which are responsible for occurrence of either E or F or both. Thus, $E \cup F$ is a new event and it occurs if and only if either E or F or both occur i.e. if and only if at least one of the events E or F occurs. Generalising this idea, we can define a new event $\bigcup_{j=1}^k E_j$, read as “union” of the k events E_1, E_2, \dots, E_k , as the event which consists of all sample points that are in at least one of the events E_1, E_2, \dots, E_k and it occurs if and only if at least one of the events E_1, E_2, \dots, E_k occurs.

Example 7: Consider Example 3 and let E_j be the event that the three tosses produce j heads, $j = 0, 1, 2, 3$. Hence, $\bigcup_{j=0}^2 E_j$ is the event that the three tosses produce at most 2 heads i.e. not all the three tosses produce heads.

* * *

Again, let E and F be two given events. We write $E \cap F$, read as E “intersection” F, to denote the collection of sample points any of whose occurrence implies the occurrence of both E and F. Thus, $E \cap F$ is a new event and it occurs if and only if both the events E and F occur. Generalising this idea, we can define a new event $\bigcap_{j=1}^k E_j$ read as “intersection” of the k events E_1, E_2, \dots, E_k , as the event which consists of sample points that are common to each of the events E_1, E_2, \dots, E_k , and it occurs only if all the k events E_1, E_2, \dots, E_k occur simultaneously.

Further, two events E and F are said to be **mutually exclusive** or **disjoint** if they do not have a common sample point i.e. $E \cap F = \phi$. Two mutually exclusive events then cannot occur simultaneously. In the coin-tossing experiment for instance, the two events, heads and tails, are mutually exclusive: if one occurs, the other cannot occur. To have a better understanding of these events let us once again look at Example 3.

Let E be the event of scoring an odd number of heads and F be the event that tail appears in the first two tosses, so that $E = \{s_1, s_5, s_6, s_7\}$ and $F = \{s_5, s_8\}$. Now $E \cap F = \{s_5\}$, the event that only the third toss yields a head. Thus events E and F are not mutually exclusive.

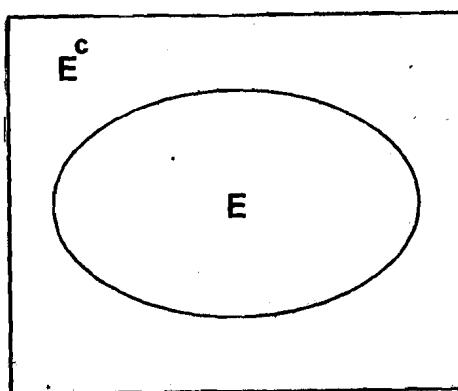


Fig. 1(a)

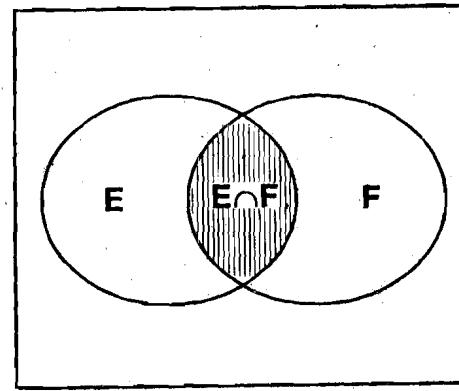


Fig. 1(b)

The above relations between events can be best viewed through a Venn diagram. A rectangle is drawn to represent the sample space ζ . All the sample points are represented within the rectangle by means of points. An event is represented by the region enclosed by a closed curve containing all the sample points leading to that event. The space inside the rectangle but outside the closed curve representing E represents the complementary event E^c (See Fig.1(a) in the previous page.) Similarly, in Fig.1(b), the space inside the curve represented by the broken line represent the event $E \cup F$ and the shaded portion represents $E \cap F$.

Why don't you try the following exercises now.

- E3) Suppose A, B and C are events in a sample space ζ . Specify whether the following relations are true or false.:
- $(A \cap B) \cap (A \cup B) = \phi$
 - $(A^c \cap B^c)^c = A \cup B$
 - $A^c \cap (A \cap B) = A \cup B^c$
 - $A \cap (B \cup B^c) = A \cap (A \cup B) = A$

After the above preliminaries, we are now ready to bring in the concept of probability of events.

2.3 PROBABILITY CONCEPTS

As is clear by now, the outcome of a random experiment being uncertain, none of the various events associated with a sample space can be predicted with certainty before the underlying experiment is performed and the outcome of it is noted. However, some events may intuitively seem to be more likely than the rest. For example, talking about human beings, the event that a person will live 20 years seems to be more likely compared to the event that the person will live 200 years. Such thoughts motivate us to explore if one can construct a scale of measurement to distinguish between likelihoods of various events. Towards this, a small but extremely significant fact comes to our help. Before we elaborate on this, we need a couple of definitions.

Consider an event E associated with a random experiment; suppose the experiment is repeated n times under identical conditions and suppose the event E (which is not likely to occur with every performance of the experiment) occurs $f_n(E)$ times in these n repetitions. Then, $f_n(E)$ is called the **frequency** of the event E in n repetitions of the experiment and $r_n(E) = f_n(E)/n$ is called the **relative frequency** of the event E in n repetitions of the experiment. Let us consider the following example.

Example 8: Consider the experiment of throwing a coin. Suppose we repeat the process of throwing a coin 5 times and suppose the following are the frequencies of a head:

No. of repetitions (n)	Frequency of head ($f_n(H)$)	Relative frequency of head $r_n(H)$
1	0	0
2	1	1/2
3	2	2/3
4	3	3/4
5	3	3/5

Notice that the third column in Table-1 gives the relative frequencies $r_n(H)$ of heads. We can keep on increasing the number of repetitions n and continue calculating the values of $r_n(H)$ in Table 1.

Merely to fix ideas regarding the concept of probability of an event, we present below a very naive approach which in no way is rigorous, but it helps to see things better at this stage.

2.3.1 Probability of an Event

While the outcome of a random experiment and hence the occurrence of any event E cannot be predicted beforehand, it is interesting to know that the relative frequency $r_n(E)$ of E though may initially fluctuate significantly, would settle down around a constant eventually i.e.

$$r_n(E) \approx r_{n+1}(E) \approx r_{n+2}(E) \approx \dots$$

for all large values of n. We shall not give a formal mathematically rigorous proof of this phenomenon at this stage but we shall try to explain it through a simple argument (we caution you at this stage that this argument is not strictly mathematically rigorous. It merely helps you to visualise that such a result is a possibility).

Observe that the frequency $f_{n+1}(E)$ of E in $n+1$ repetitions must be equal to either $f_n(E)$ or $f_n(E) + 1$ depending on whether E did not or did occur at the $(n+1)$ -th repetition. Thus,

$$\begin{aligned} r_{n+1}(E) - r_n(E) &= f_{n+1}(E)/(n+1) - f_n(E)/n \\ &= \{nf_{n+1}(E) - (n+1)f_n(E)\} / \{(n+1)n\} \\ &= \{nf_n(E) - (n+1)f_n(E)\} / \{(n+1)n\}, \text{ or} \\ &\quad \{n(f_n(E) + 1) - (n+1)f_n(E)\} / \{(n+1)n\} \\ &= -f_n(E)/n(n+1), \text{ or} \\ &\quad (n - f_n(E))/(n+1)n, \end{aligned}$$

so that

$$\begin{aligned} |r_{n+1}(E) - r_n(E)| &= \left| -\frac{f_n(E)/n}{n+1} \right| \leq \frac{1}{n+1}, \text{ or} \\ &= \left| \frac{1 - f_n(E)/n}{n+1} \right| \leq \frac{1}{n+1} \end{aligned}$$

as $f_n(E)/n$ being the relative frequency of E in n repetitions can never exceed 1. Thus, the difference between $r_{n+1}(E)$ and $r_n(E)$ can be made as small as we please by increasing the number of repetitions n.

In any case, the constant around which the relative frequency of an event E settles down as the number of repetitions becomes large (i.e. $\lim_{n \rightarrow \infty} r_n(E)$) is called the **probability** of E and is denoted as $P(E)$. Thus, $P(E)$ can be interpreted to be the proportion of times one would expect the event E to take place when the underlying random experiment is repeated a large number of times under identical conditions. How large should the number of repetitions be in order for the relative frequency to settle down is another matter. Let us now illustrate through examples what we have discussed above.

Example 9: When we say that the probability of scoring a head when it is tossed is 0.5, we merely mean that 50% of a **large number of tosses** should result in heads. We do not mean that in 10 tosses, 5 will turn heads and 5 tails. However if the coin is tossed N times and N is sufficiently large, then we may expect nearly $N/2$ heads and $N/2$ tails.

* * *

Example 10: When we say that the probability of a certain machine failing before 500 hours of operation is 2%, we merely mean that out of a large number of such machines, about 2% of them will fail before 500 hours of operation.

* * *

Evidently, then such a numerical assessment of likelihood of the event in a given situation helps make an appropriate decision. For example, suppose we have two

brands A and B of an equipment available in the market such that the probability of failure before 500 hours of operation is 0.02 for an equipment of brand A and 0.10 for an equipment of brand B. We will obviously choose the former variety as the long term proportion of these failing before 500 hours of operation is less.

Notice that the relative frequency of any event E in n repetitions, $r_n(E)$, satisfies the inequalities $0 \leq r_n(E) \leq 1$ and therefore $P(E)$ the limit of $r_n(E)$ as $n \rightarrow \infty$, must satisfy

$$0 \leq (P(E)) \leq 1 \quad (1)$$

Further, since ζ denotes the event that the experiment will result into one of the outcomes listed in ζ and since by definition, it consists of all possible outcomes, $r_n(\zeta) = 1$, for every n, so that

$$P(\zeta) = 1 \quad (2)$$

Consider two **mutually exclusive** events E and F. Then, to calculate $r_n(E \cup F)$, i.e., the proportion of times either E or F is observed, we must add the proportion of times E is observed and the proportion of times F is observed so that $r_n(E \cup F) = r_n(E) + r_n(F)$, and hence in the limit,

$$P(E \cup F) = P(E) + P(F), \quad (3)$$

Remember that E and F cannot occur together as they are mutually exclusive. Arguing in a similar manner, one observes that if E_1, E_2, \dots, E_k are k mutually exclusive events, then

$$P(\bigcup_{j=1}^k E_j) = P(E_1) + P(E_2) + \dots + P(E_k) \quad (4)$$

i.e., the probability of observing at least one of the k mutually exclusive events E_1, E_2, \dots, E_k (and hence exactly one) is equal to the sum of their individual probabilities.

From the above properties (1)-(4) of probability, the following useful results follow immediately:

PROPOSITION 1: In view of relations (1) and (2), for any event E,

$$1 = P(\zeta) = P(E \cup E^c)$$

so that for any event E,

$$P(E^c) = 1 - P(E). \quad (5)$$

You will realise later that this result is particularly useful in many problems where it is easier to compute $P(A^c)$ than $P(A)$. So whenever we wish to evaluate $P(A)$, we compute $P(A^c)$ and get the desired result by subtraction from unity

PROPOSITION 2: In Eqn.(5) above, substituting ζ for E, so that $\zeta^c = \phi$, we obtain

$$P(\phi) = 1 - P(\zeta) = 1 - 1 = 0 \quad (6)$$

Notice that you may come across situations where the converse of result (6) is not true. That is, if $P(A) = 0$, we cannot in general conclude that $A = \phi$, for, there are situations in which we assign probability zero to an event that can occur.

PROPOSITION 3: Consider a finite sample space $\zeta = \{s_1, s_2, \dots, s_m\}$. Suppose, from the nature of the experiment, it is reasonable to assume that all the possible outcomes s_1, s_2, \dots, s_m are equally likely i.e. after taking into consideration all the relevant information pertaining to the experiment, none of the m outcomes seems to be more likely than the rest. For $i = 1, 2, \dots, m$, let E_i be the simple event defined as

$$E_i = \{s_i\},$$

Since, all these simple events are equally likely,

$$P(E_1) = P(E_2) = \dots = P(E_m) = p, \text{ say}$$

Now,

$$1 = P(\zeta) = P(\bigcup_{j=1}^m E_j) = P(E_1) + P(E_2) + P(E_3) + \dots + P(E_m)$$

so that for $i = 1, 2, \dots, m$,

$$P(E_i) = p = 1/m$$

Also, any event E that consists of k of the m sample points ($k \leq m$) can be represented as the union of the k simple events (which are trivially mutually exclusive). Arguing as above, we conclude that

$$P(E) = kp \quad (7)$$

$$\begin{aligned} &= k/m \\ &= \frac{(\text{Total no. of sample points in } E)}{(\text{Total no. of sample points in } \zeta)} \\ &= \frac{(\text{Total no. of outcomes favouring } E)}{(\text{Total no. of the outcomes of the experiment})} \\ &= \left. \frac{(\text{Total no. of ways } E \text{ can occur})}{(\text{Total no. of the outcomes of the experiment.})} \right\} \end{aligned} \quad (8)$$

k events are mutually exclusive if no two of them have any element in common

Example 11: Let us once again consider Example 3. Whenever the coin is unbiased, H and T are equally likely and then all the eight simple events are equally likely. Then for an event E that the three tosses produce at least one head, Formula (8) gives $P(E) = 7/8$, since E consists of 7 sample points and the sample space S consists of 8 sample points.

* * *

Example 12: Suppose a fair die is rolled once and the score on the top face is recorded. Obviously, the sample space $\zeta = \{1, 2, 3, 4, 5, 6\}$. Since the die is fair, all the six outcomes are equally probable so that the probability of scoring i is $1/6$, for $i = 1, 2, \dots, 6$. Now, consider the event E that the top face shows an odd score; clearly, $E = \{1, 3, 5\}$. Thus, $P(E) = 3/6 = 1/2$.

* * *

Example 13: Consider two fair dice which are rolled simultaneously and the scores are recorded. Hence, the sample space consisting of $6 \times 6 = 36$ points is

$$\begin{aligned} \zeta &= \{(1, 1), (1, 2), \dots, (6, 6)\} \\ &= \{(i, j) : i = 1, 2, \dots, 6; j = 1, 2, \dots, 6\} \end{aligned}$$

Here, a typical sample point is denoted by (i, j) where i is the score on the first die and j is the score on the second die. Now, consider the event E that the sum of the two scores is 7 or more. Then,

$E = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), (2, 6), (3, 5), (4, 4), (5, 3), (6, 2), (3, 6), (4, 5), (5, 4), (6, 3), (4, 6), (5, 5), (6, 4), (5, 6), (6, 5), (6, 6)\}$. Since the dice are fair, all the 36 outcomes listed in ζ are equally likely and since E comprises of 21 sample points, $P(E) = 21/36 = 7/12$.

Now consider the event F that none of the two scores is even. To calculate the probability of F , we can surely list out all the outcomes, which favour this event. However, to economise on computations, we may like to use a simple argument as follows. Since none of the scores is even, then, both must be odd. Now, on each die, the possible odd scores are 1, 3 or 5. Also with each odd score on the first die, any of the three possible odd scores on the second die may be associated. Thus with each odd score on the first die, there will then be three possible results of the complete experiment and since there are 3 possible odd scores on the first die, the total number of sample points in F is $3 \times 3 = 9$. Thus, $P(F) = 9/36 = 1/4$.

Let us consider another event F' described by the condition that the only possible scores on the two dice is either 1 or 3. Then the number of possible sample points in F' are $2 \times 2 = 4$ and $P(F') = 4/36 = 1/9$.

You may notice here that $E \cap F' = \emptyset$, since the sum of the two scores corresponding to any sample point in F' can be at most 6 so that $P(E \cap F') = 0$.

* * *

Example 14: Consider the case of drawing marbles without replacement as in

Example 6. Suppose, we are interested in working out the probability of the event E that the two marbles drawn are of different colour. When we are reaching out for marbles blindly and the marbles are of identical size, we are not favouring any particular marble so that all the possible pairs are equally likely to be picked up. Now, the total number of sample points in S is 6 and the number of sample points favouring the occurrence of the event E is 4, so that $P(E) = 4/6 = 2/3$.

And now some exercises for you

E4) In Example 14 above find the probability of the event E when drawing is done, with replacement.

E5) The manager of a shoe store sells from 0 to 4 pairs of shoes of a recently introduced design every day. Based on experience, the following probabilities are assigned to daily sales of 0, 1, 2, 3, or 4 pairs:

$$\begin{array}{rcl} P(0) & = & 0.08 \\ P(1) & = & 0.18 \\ P(2) & = & 0.32 \\ P(3) & = & 0.30 \\ P(4) & = & \frac{0.12}{1.00} \end{array}$$

- a) Are these valid probability assignments? Why or why not?
- b) Let A be the event that 2 or fewer are sold in a day. Find $P(A)$.
- c) Let B be the event that 3 or more are sold in a day. Find $P(B)$

Before proceeding further we have the following remarks

Remark 1: For a population of N objects a_1, a_2, \dots, a_N , choosing one object at random from these N objects would mean that the selection mechanism is such that each of the N objects has probability $1/N$ of being chosen.

Remark 2: For a population of N objects a_1, a_2, \dots, a_N , choosing n ($n \leq N$) objects at random from these N objects would mean that all the combinations of n objects out of N objects are equally likely. This can be achieved provided the selection mechanism is such that objects are selected one after the other and at each draw all the objects still remaining in the population are equally likely to be chosen.

Let us now take up some more examples to illustrate the use of Formula (8) for calculating probabilities in the cases of selections with/without replacement.

Problem 1: Two balls are drawn at random from a bowl containing 6 balls, of which four balls are white and two are red. What is the probability that (a) both balls are white, (b) both balls are of the same colour, and (c) at least one of the balls is white?

Solution: Case 1: Balls drawn with replacement: Let us number the balls as 1, 2, 3, 4, 5 and 6 such that the balls numbered as 1-4 are white while the balls numbered as 5 and 6 are red. We are required to draw 2 balls randomly with replacement. For the first draw, we have 6 choices; since the ball chosen in the first draw is sent back to the bowl before the second ball is drawn, for the second draw also, we have 6 choices. Thus, the sample space $\zeta = \{(i, j) | i = 1, 2, \dots, 6; j = 1, 2, \dots, 6\}$ has $6 \times 6 = 36$ sample points. Let E_j be the event that the j of the two balls selected are white, $j = 0, 1, 2$. Then

- a) The required probability of both the balls being white will be $P(E_2)$. In order for E_2 to take place, the outcome (i, j) must be such that both i and j are numbers between 1 and 4. Since the ball chosen in the first draw is sent back to the bowl before the second ball is drawn, the number of sample points belonging to E_2 is $4 \times 4 = 16$. Hence $P(E_2) = 16/36 = 4/9$.

- b) In this case the event we are interested in is $E_2 \cup E_0$ because E_2 is the event that both selected balls are white while E_0 is the event that none of the two selected ball is white implying that both must be red. Then the required probability is $P(E_2 \cup E_0) = P(E_2) + P(E_0)$, since these two events are mutually exclusive. In view of an argument of the previous case, the number of sample points belonging to E_0 is $2 \times 2 = 4$. Hence the required probability is $16/36 + 4/36 = 20/36 = 5/9$.

- c) Now the event that at least one of the two balls drawn is white is complementary to E_0 . Thus the probability that at least one of the two balls drawn is white is $1 - P(E_0) = 1 - 4/36 = 32/36 = 8/9$.

We shall now consider the case of finding probabilities without replacement.

Case 2: Balls drawn without replacement: Here, we are required to draw 2 balls randomly without replacement. For the first draw, we have 6 choices; since the ball chosen in the first draw is not being sent back to the bowl before the second ball is drawn, for the second draw, we have 5 choices. Thus, the sample space

$\zeta = \{(i, j) | i = 1, 2, \dots, 6; j = 1, 2, \dots, 6; i \neq j\}$ has $6 \times 5 = 30$ sample points (combinations of 2 out of 6 balls, keeping in mind the order in which the balls are drawn). Let E_j be the event that the j of the two balls selected are white, $j=0,1,2$.

- a) The required probability will be $P(E_2)$. In order for E_2 to take place, the outcome (i, j) must be such that both i and j are numbers between 1 and 4. Since the ball chosen in the first draw is not sent back to the bowl before the second ball is drawn, the number of sample points belonging to E_2 is $4 \times 3 = 12$. Hence $P(E_2) = 12/30 = 2/5$.
- b) Here the event we are interested in is $E_2 \cup E_0$ and as such the required probability is $P(E_2 \cup E_0) = P(E_2) + P(E_0)$, since these two events are mutually exclusive. In view of an argument of the previous case, the number of sample points belonging to E_0 is $2 \times 1 = 2$. Hence the required probability is $12/30 + 2/30 = 14/30 = 7/15$.
- c) The event that at least one of the two balls drawn is white is complementary to E_0 . Thus the probability that at least one of the two balls drawn is white is $1 - P(E_0) = 1 - 2/30 = 28/30 = 14/15$.

————— X —————

Problem 2: Consider a workshop that employs three mechanics in shifts to repair the machines as and when these fail. During a given shift, 4 machines failed and the repair duties were assigned to the three mechanics through a system of lottery such that for each failed machine, all the three mechanics were equally likely to be chosen. What is the probability that none of the mechanics as well as machine remained idle in that shift?

Solution: Suppose we name the mechanics as 1,2 and 3 and designate the four machines as I, II, III and IV. The result of an allocation of mechanics to the four machines can be represented as (i,j,k,l) meaning that mechanic i is assigned to machine I, mechanic j is assigned to machine II, mechanic k is assigned to machine III and mechanic l is assigned to machine IV, where i,j,k and l are each numbers between 1 and 3 designating the concerned mechanic. Note that there being only three mechanics and four machines, i,j,k,l cannot be all different.

Thus, $\zeta = \{(i, j, k, l) | i = 1, 2, 3; j = 1, 2, 3; k = 1, 2, 3; l = 1, 2, 3\}$. Since the allocation of the same mechanic to more than one machine is feasible and conversely one or two mechanics being idle is also feasible, each possible value of i can be combined with each possible value of j,k and l and since each of i,j,k and l can adopt three values, the total number of sample points in ζ must be $3 \times 3 \times 3 \times 3 = 3^4 = 81$.

Now if all the machines have to have work on that day then one of the mechanics must have been assigned two machines and the other two mechanics one each. This is the

only way the event "no machine and no mechanics is idle" can occur. In terms of the sample point (i,j,k,l) this would mean that two of the numbers i,j, k and l must be equal and the other two must be different and different from each other. For example, a typical allocation leading to all the mechanics being engaged and no machine is idle on that day can be (1,1,2,3). To count all the possible allocations so that everybody is engaged, we have to choose two of the indices i,j,k and l that are to be identical and this can be done in ${}^4C_2 = 6$ ways. While these two indices are identical the other two indices can be switched around in 2 ways. Thus, for example, if we allocate machines I and II to mechanic 1, then machines III and IV can be allocated to mechanics 2 and 3 respectively or to mechanics 3 and 2 respectively. Since there are altogether 3 mechanics, the total number of ways the machines can be allocated to them so that all the three mechanics are engaged and no machine is idle will be $3 \times {}^4C_2 \times 2 = 36$. Thus the required probability = $36/81 = 4/9$.

X

Why don't you try the following exercises now?

- E6) Suppose that in a library, two of the six copies of a book are damaged. If the library assistant selects 2 out of the six copies at random, what is the probability that he will select (i) the two damaged copies? (ii) at least one of the two damaged copies?
- E7) A student takes a multiple-choice test composed of 100 questions, each with six possible answers. If, for each question, he rolls a fair die to determine the answer to be marked, what is the probability that he answers 20 questions rightly?

We now discuss the rule for finding the probability of the union of any two events, disjoint or not.

2.3.2 Probability of Compound Events

We have seen that for two mutually exclusive events E and F, $P(E \cup F) = P(E) + P(F)$. But this formula cannot be used, for example, to find the probability that at least one of two friends will pass a language examination or the probability that a customer will buy a shirt, a sweater, a belt, or a tie at a departmental store. Both friends can pass the examination and customer of the departmental store can buy any number of these items. To find a formula for $P(E \cup F)$ which holds regardless of whether E and F are mutually exclusive or not, let us consider the following proposition

PROPOSITION 4: For any two events E and F,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F). \quad (9)$$

Let us draw the Venn diagram as in Fig.2.

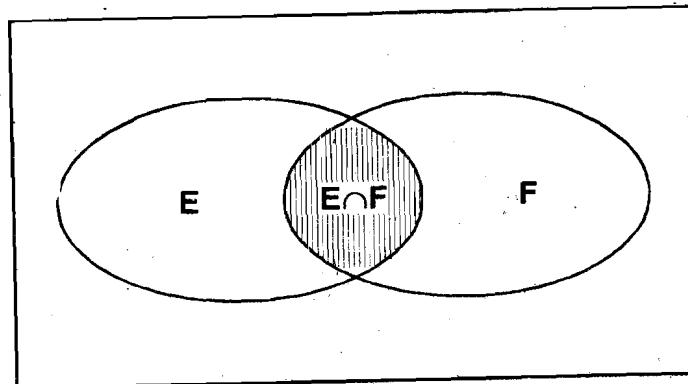


Fig.2

Note that each of the events E and F can be decomposed into mutually exclusive components as

$$E = (E \cap F^c) \cup (E \cap F)$$

$$F = (E^c \cap F) \cup (E \cap F)$$

Then by Formula (3)

$$P(E) = P(E \cap F^c) + P(E \cap F) \quad (10)$$

$$P(F) = P(E^c \cap F) + P(E \cap F) \quad (11)$$

Hence, by summing the two Eqns. (10) and (11), we have

$$\begin{aligned} P(E) + P(F) &= \{P(E \cap F^c) + P(E \cap F) + P(E^c \cap F)\} + P(E \cap F); \\ &= P(E \cup F) + P(E \cap F), \end{aligned} \quad (12)$$

since $(E \cap F^c) \cup (E \cap F) \cup (E^c \cap F) = E \cup F$, our result (9) follows from Eqn.(12).

Observe that the above result reduces to Formula 3 whenever E and F are mutually exclusive; this is so because if E and F are mutually exclusive, then $E \cap F = \phi$ and $P(\phi) = 0$.

PROPOSITION 5: For any k events $E_i, i = 1, 2, \dots, k$, the above formula can be further generalised for calculation of the probability of occurrence of at least one of these events as follows:

$$P(\bigcup_{i=0}^k E_i) = S_1 - S_2 + S_3 - \dots + (-1)^{k-1} S_k \quad (13)$$

where,

$$S_1 = P(E_1) + P(E_2) + \dots + P(E_k)$$

$$\begin{aligned} S_2 &= P(E_1 \cap E_2) + P(E_1 \cap E_3) + P(E_1 \cap E_4) + \dots + P(E_2 \cap E_3) \\ &\quad + P(E_2 \cap E_4) + \dots + P(E_{k-1} \cap E_k), \\ &\quad \dots + P(E_{k-2} \cap E_{k-1} \cap E_k), \end{aligned}$$

⋮ ⋮ ⋮

$$S_k = P(E_1 \cap E_2 \cap E_3 \cap \dots \cap E_k).$$

For example, whenever k = 4, Formula (13) will be as follows:

$$P(\bigcup_{i=0}^4 E_i) = S_1 - S_2 + S_3 - S_4$$

where,

$$S_1 = P(E_1) + P(E_2) + P(E_3) + P(E_4)$$

$$\begin{aligned} S_2 &= P(E_1 \cap E_2) + P(E_1 \cap E_3) \\ &\quad + P(E_1 \cap E_4) + P(E_2 \cap E_3) + P(E_2 \cap E_4) + P(E_3 \cap E_4), \end{aligned}$$

$$S_3 = P(E_1 \cap E_2 \cap E_3) + (E_1 \cap E_2 \cap E_4) + P(E_2 \cap E_3 \cap E_4),$$

$$S_4 = P(E_1 \cap E_2 \cap E_3 \cap E_4).$$

Given Proposition 4, Formula (13) can be proved by induction. We, however, do not discuss the proof here but shall illustrate it through a problem.

Problem 3: Suppose that a candidate has applied for admission to three management schools A,B and C. It is known that his chances of being selected by A,B,C,A and B both, A and C both, B and C both are 0.47, 0.29, 0.22, 0.08, 0.06, 0.07 respectively. It is also known that the chance of his being selected by all the three schools is 0.03. What is the probability that he will be selected by none?

Solution: Suppose A represents the event that he will be selected by the school A; events B and C are similarly defined. From above, we know that

$$P(A) = 0.47, P(B) = 0.29, P(C) = 0.22, P(A \cap B) = 0.08,$$

$$P(A \cap C) = 0.06, P(B \cap C) = 0.07, P(A \cap B \cap C) = 0.03. \text{ Hence,}$$

$$S_1 = P(A) + P(B) + P(C) = 0.47 + 0.29 + 0.22 = 0.98;$$

$$S_2 = P(A \cap B) + P(A \cap C) + P(B \cap C) = 0.08 + 0.06 + 0.07 = 0.21; S_3 = 0.03.$$

Thus, as per Proposition 5, the probability that he will be selected by at least one of the

schools A, B and C will be $P(A \cup B \cup C) = S_1 - S_2 + S_3 = 0.98 - 0.21 + 0.03 = 0.80$. Since the event that he will be selected by none is complementary to the event that he will be selected by at least one of the schools A, B and C i.e. $A \cup B \cup C$, the required probability $P(A^c \cap B^c \cap C^c) = 1 - P(A \cup B \cup C) = 1 - 0.80 = 0.20$.

Based on Proposition 5, we may further develop an inequality, called Boole's Inequality which provides an upper bound for the probability of occurrence of at least one out of k events in terms of the probabilities of these individual events. This inequality obviously provides a way to check probability computations in a situation where it would be computationally involved to apply Formula 13.

PROPOSITION 6: For any k events $E_i, i = 1, 2, \dots, k$ the above formula can be further generalised for calculating the probability of occurrence of at least one of these events as follows:

$$P(\bigcup_{i=1}^k E_i) \leq P(E_1) + P(E_2) + \dots + P(E_k) \quad (14)$$

Let us define $B_j = \bigcup_{i=1}^j E_i, j = 1, 2, \dots, k$. Then,

$$\begin{aligned} P(\bigcup_{i=1}^k E_i) &= P(B_k) = P(B_{k-1} \cup E_k) \\ &= P(B_{k-1}) + P(E_k) - P(B_{k-1} \cap E_k) \\ &\leq P(B_{k-1}) + P(E_k) \quad (\text{since probability of any event is non-negative}) \\ &\leq P(B_{k-2}) + P(E_{k-1}) + P(E_k) \\ &\leq P(B_{k-3}) + P(E_{k-2}) + P(E_{k-1}) + P(E_k) \\ &\vdots \\ &\leq P(E_k) + P(E_2) + \dots + P(E_{k-2}) + P(E_{k-1}) + P(E_k) \end{aligned}$$

And now a few exercises for you.

-
- E8) Two dice are thrown n times simultaneously. Find the probability that each of the six combinations $(1, 1), (2, 2), \dots, (6, 6)$ appears at least once?
- E9) How many times should an unbiased coin be tossed in order that the probability of observing at least one head be equal to or greater than 0.9?
-

We now introduce another concept which is important in the study of probability theory.

2.4 CONDITIONAL PROBABILITY AND INDEPENDENT EVENTS

Let ζ be the sample space corresponding to an experiment and E and F are two events of ζ . Suppose the experiment is performed and the outcome is known only partially to the effect that the event F has taken place. Thus there still remains a scope for speculation about the occurrence of the other event E . Keeping this additional piece of information confirming the occurrence of F in view, it would be appropriate to modify the probability of occurrence of E suitably. That such modifications would be necessary can be readily appreciated through two simple instances as follows:

Example 15: Suppose, E and F are such that $F \subset E$ so that occurrence of F would automatically imply the occurrence of E . Thus with the information that the event F has taken place in view, it is plausible to assign probability 1 to the occurrence of E irrespective of its original probability.

Example 16: Suppose, E and F are two mutually exclusive events and thus they cannot occur together. Thus whenever we come to know that the event F has taken place, we can rule out the occurrence of E. Therefore, in such a situation, it will be appropriate to assign probability 0 to the occurrence of E.

* * *

Example 17: Suppose a pair of balanced dice A and B are rolled simultaneously so that each of the 36 possible outcomes is equally likely to occur and hence has probability $\frac{1}{36}$. Let E be the event that the sum of the two scores is 10 or more and F be the event that exactly one of the two scores is 5.

Then $E = \{(4, 6), (5, 5), (5, 6), (6, 4), (6, 5), (6, 6)\}$ so that $P(E) = 6/36 = 1/6$.

Also, $F = \{(1, 5), (2, 5), (3, 5), (4, 5), (6, 5), (5, 1), (5, 2), (5, 3), (5, 4), (5, 6)\}$.

Now suppose we are told that the event F has taken place (note that this is only partial information relating to the outcome of the experiment). Since each of the outcome originally had the same probability of occurring, they should still have equal probabilities. Thus given that exactly one of the two scores is 5 each of the 10 outcomes of event F has probability $\frac{1}{10}$, while the probability of remaining 26 points in the sample space is 0. In the light of the information that the event F has taken place the sample points $(4, 6), (6, 4), (5, 5)$ and $(6, 6)$ in the event E must not have materialised. One of the two sample points $(5, 6)$ or $(6, 5)$ must have materialised. Therefore probability of E would no longer be $1/6$. Since all the 10 sample points in F are equally likely, the revised probability of E which given the occurrence of F, which occur through the materialization of one of the two sample points $(6, 5)$ or $(5, 6)$ should be $2/10 = 1/5$.

* * *

The probability just obtained is called the conditional probability that E occurs given that F has occurred and is denoted by $P(E|F)$. We shall now derive a general formula for calculating $P(E|F)$.

Consider the following probability table:

Table 2

Events	E	E^c
F	p	q
F^c	r	s

In Table 2, $P(E \cap F) = p$, $P(E^c \cap F) = q$, $P(E \cap F^c) = r$ and $P(E^c \cap F^c) = s$ and hence, $P(E) = P((E \cap F) \cup (E \cap F^c)) = P(E \cap F) + P(E \cap F^c) = p + r$ and similarly, $P(F) = q + s$.

Now suppose that the underlying random experiment is being repeated a large number of times, say N times. Thus, taking a cue from the long term relative frequency interpretation of probability, the approximate number of times the event F is expected to take place will be $NP(F) = N(q + s)$. **Under the condition that the event F has taken place**, the number of times the event E is expected to take place would be $NP(E \cap F)$ as both E and F must occur simultaneously. Thus, the long term relative frequency of E under the condition of occurrence of F, i.e. the probability of occurrence of E under the condition of occurrence of F, should be $NP(E \cap F)/NP(F) = P(E \cap F)/P(F)$. This is the proportion of times E occurs out of the repetitions where F takes place.

With the above background, we are now ready to define formally the conditional probability of an event given another.

Definition: Let E and F be two events from a sample space ζ . The **conditional probability** of the event E given the event F, denoted by $P(E|F)$, is defined as

$$P(E|F) = P(E \cap F)/P(F), \text{ whenever } P(F) > 0. \quad (15)$$

When $P(F) = 0$, we say that $P(E|F)$ is undefined. We can also write from Eqn.(15)

$$P(E \cap F) = P(E|F)P(F). \quad (16)$$

Referring back to Example 17, we see that $P(E) = 6/36$, $P(F) = 10/36$; since, $E \cap F = \{(5, 6), (6, 5)\}$, $P(E \cap F) = 2/36$.

From Result (15), $P(E|F) = (2/36)/(10/36) = 2/10 = 1/5$, which is same as that obtained in Example 17.

Result (16) can be generalised to k events E_1, E_2, \dots, E_k , where $k \geq 2$. And now an exercise for you.

E10) Two fair dice are rolled simultaneously. What is the conditional probability that the sum of the scores on the two dice will be 7 given that (i) the sum is odd (ii) the sum is greater than 6, (iii) the outcome of the first die was odd, (iv) the outcome of the second die was even (v) the outcome of at least one of the dice was odd?

Let us now consider some more applications of Formula (16).

Problem 4: A blood disease is present in 12% of a population and is not present in the remaining 88%. An imperfect clinical test successfully detects the disease and with probability 0.90. Thus, if a person has the disease in the serious form, the probability is 0.9 that the test will be positive and it is 0.1, if the test is negative. Moreover, among the unaffected persons, the probability that the test will be positive is 0.05.

- A person selected at random from the population is given the test and the result is positive. What is the probability that this person has the disease?
- What is the probability that the test correctly detects the disease?

Solution: Let E be the event that a person has the disease and F be the event that the blood test is positive. From the given data, we note that

$$P(E) = 0.12, P(F|E) = 0.90, P(F^c|E) = 0.10, P(F|E^c) = 0.05$$

- We are required to compute $P(E|F)$.

$$\text{By definition, } P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

$$\text{We have, } P(E \cap F) = P(F|E)P(E) = (0.90)(0.12) = 0.108$$

Also,

$$\begin{aligned} P(F) &= P(F \cap E) + P(F \cap E^c) \\ &= P(F|E)P(E) + P(F|E^c)P(E^c) \\ &= (0.90)(0.12) + (0.05)(0.88) \\ &= 0.108 + 0.044 \\ &= 0.152 \end{aligned}$$

$$\text{Hence the required conditional probability } P(E|F) = 0.108/0.152 = 0.7105.$$

- Let G be the event that the test correctly detects the disease. Then G will consist of all those who actually have the disease and their blood test is positive and also those who do not have the disease and their blood test is negative.

Then we can write $G = (E \cap F) \cup (E^c \cap F^c)$

so that $P(G) = P(E \cap F) + P(E^c \cap F^c)$

$$\text{Now } P(E^c \cap F^c) = P(F^c|E^c)P(E^c) = \{1 - P(F|E^c)\} P(E^c) = (0.95)(0.88) = 0.836$$

$$\text{Thus } P(G) = 0.108 + 0.836 = 0.944$$

————— X —————

Before we go further you may try these exercises.

E11) Consider a family with two children. Assume that each child is as likely to be a boy as it is to be a girl. What is the conditional probability that both children are boys given that (i) the elder child is a boy (ii) at least one of the children is a boy?

From the definition of conditional probability, you may observe that whenever E and F are two mutually exclusive events then $P(E|F) = 0$; similarly, if $F \subset E$, then $P(E|F) = 1$. In each of these cases, the knowledge of occurrence of F gives us a definite idea about the probability of occurrence of E. However, there are many situations where the knowledge of occurrence of some event F hardly have any bearing whatsoever on the occurrence or non-occurrence of another event E. To understand this let us once again consider Problem 1.

Suppose E is the event that the first ball drawn is white and F be the event that the second ball drawn is red. Let us consider the following two cases:

Case 1: Drawing balls with replacement: Arguing as in Problem 1, $P(E \cap F) = (4 \times 2)/36$, $P(E) = (4 \times 6)/36 = 4/6$ and $P(F) = (6 \times 2)/36 = 2/6$. Hence, $P(E|F) = P(E \cap F)/P(F) = 4/6$ which is the same as the unconditional probability of E.

In other words, the occurrence of F had no influence on the occurrence or non-occurrence of E. In this sense, therefore, the two events E and F are **independent**.

Case 2: Drawing balls without replacement: Arguing as in Problem 1, $P(E \cap F) = (4 \times 2)/30$, $P(E) = (4 \times 5)/30 = 4/6$ and $P(F) = (5 \times 2)/30 = 2/6$. Hence, $P(E|F) = P(E \cap F)/P(F) = 2/5$ which is different from the unconditional probability of E.

In other words, the occurrence of F had an influence on the occurrence or non-occurrence of E. In this sense, therefore, the two events E and F are **dependent**.

Definition Two events E and F from a given sample space are said to be **independent** if and only if

$$P(E \cap F) = P(E)P(F);$$

otherwise, they are said to be **dependent**.

Again, using the definition of conditional probability, one can equivalently say that two events E and F from a given sample space are independent if and only if

$$P(E|F) = P(E), \text{ and}$$

$$P(F|E) = P(F).$$

Consider the following example.

Example 18: Suppose an unbiased coin is tossed twice. Let F be the event that the first toss results in a head and E be the event that the second toss produces a head. The sample space in this case is $\zeta = \{(H, H), (H, T), (T, H), (T, T)\}$ and the coin being unbiased, all these outcomes are equally likely so that each has probability 0.25.

Here, $E = \{(H, H), (T, H)\}$ and $F = \{(H, H), (H, T)\}$ so that $E \cap F = \{(H, H)\}$.

As such, $P(E) = 2/4 = 0.50$, $P(F) = 2/4 = 0.50$, $P(E \cap F) = 0.25$.

Thus, $P(E \cap F) = 0.25 = (0.50)(0.50) = P(E)P(F)$.

Hence E and F are independent events.

* * *

And now some exercises for you.

E12) Kalpana (K) and Rahul (R) are taking a statistics course which has only 3 grades A, B and C. The probability that K gets a B is 0.3, the probability that R gets a B is 0.4. The probability that neither gets an A, but at least one gets a B is .42. The probability that K gets an A and R gets a B is 0.08. What is the probability that atleast one gets a B but neither gets a C, (assume the grades of the two students to be independent).

E13) The probability that at least one of two independent events occur is 0.5.

Probability that the first event occurs but not the second is $3/25$. Also the probability that the second event occurs but not the first is $8/25$. Find the probability that none of the two events occur.

- E14) Suppose that A and B are independent events associated with an experiment. If the probability that A or B occurs equals 0.6 while the probability that A occurs equals 0.4, determine the probability that B occurs.
-

We now end this unit by giving a summary of what we have covered in it.

2.5 SUMMARY

In this unit we have covered the following.

- 1) Experiments whose outcomes cannot be precisely predicted primarily because the totality of the factors influencing the outcomes are either not identifiable or not controllable at the time of experimentation, even if known, are called **random experiments**.
 - 2) Random experiment may comprise of a series of smaller sub-experiments called **trials**
 - 3) Each outcome of an experiment is a **sample point** and a set of all possible sample points constitute the **sample space** of the experiment.
 - 4) A specific collection or subset of sample points is called an **event**.
 - 5) Two events E and F are **mutually exclusive** or disjoint if they do not have a common sample point i.e. $E \cup F = \phi$. Two mutually exclusive events cannot occur simultaneously.
 - 6) If an experiment is repeated n number of times under identical conditions and an event E occurs $f_n(E)$ times in these n repetitions then $f_n(E)/n$ is the **relative frequency** of the event E in n repetitions of the experiment.
 - 7) The **probability** $P(E)$ of the event E is the proportion of times an event E takes place when the underlying random experiment is repeated a large number of times under identical conditions.
 - 8) For two events E and F, the probability of an event E under the condition that F has already occurred is the **conditional probability** $P(E|F)$ of E.
 - 9) Two events for which the occurrence of one has no influence on the occurrence or non-occurrence of the other are **independent** events, otherwise, they are **dependent**.
-

2.6 SOLUTIONS/ANSWERS

- E1) a) $\zeta = \{g, d\}$, where g stands for “good” or “non-defective” and d stands for “defective”
- b) Suppose we code the six colours by the numerals 1, 2, 3, 4, 5 and 6 and only one ball is drawn. Hence $\zeta = \{1, 2, 3, 4, 5, 6\}$.
- c) Let the six categories be coded as 1, 2, 3, 4, 5 and 6. Here, $\zeta = \{1, 2, 3, 4, 5, 6\}$.
- d) Suppose we code the ten cards by the numerals 1, 2, ..., 10.
 $\zeta = \{(x, y) | 1 \leq x \leq 10, 1 \leq y \leq 10x, y \text{ integers}\}$
- E2) Suppose i is the result on the first throw and j is the result on the second throw, $i = 1, 2, \dots, 6$ and $j = 1, 2, \dots, 6$. Thus, the sample space is as follows:
 $S = \{(i, j) : i = 1, 2, \dots, 6 \text{ and } j = 1, 2, \dots, 6\}$.

- (i) Event that the maximum score is 6
 $= \{(6, j) : j = 1, 2, \dots, 6\} \cup \{(i, 6) : i = 1, 2, \dots, 6\}.$
- (ii) Event the total score is 9
 $= \{(i, j) : i = 1, 2, \dots, 6, j = 1, 2, \dots, 6 \text{ and } i + j = 9\}$
- (iii) Event that each throw results in an even score
 $= \{(i, j) : i = 2, 4, 6 \text{ and } j = 2, 4, 6\}$
- (iv) Event that each throw results in an even score larger than 2
 $= \{(i, j) : i = 4, 6 \text{ and } j = 4, 6\}$
- (v) Event that the scores on the two throws differ by at least 2
 $= \{(i, j) : i = 1, 2, \dots, 6; j = 1, 2, \dots, 6 \text{ and } |i - j| > 2\}.$

- E3) (a) Consider $S = \{i : i = 1, 2, \dots, 6\}$
 Take $A = \{1, 2, 3, 4\}, B = \{3, 4, 5, 6\}$. Then
 $(A \cap B) \cap (A \cup B) = \{3, 4\} \cap \{1, 2, 3, 4, 5, 6\} = \{3, 4\} \neq \emptyset$
 so that the statement is false.
- (b) $\omega \in (A^c \cap B^c)$ if and only if $\omega \notin (A^c \cap B^c)^c$
 ie if and only if ω is neither in A^c nor in B^c
 i.e. if and only if ω is in A or in B
 i.e. if and only if $\omega \in A \cup B$. Thus, the statement is true.
- (c) Take the example of (a) above.
 $A^c \cap (A \cap B) = \{5, 6\} \cap \{3, 4\} = \emptyset \neq \{1, 2, 3, 4\} \cup \{1, 2\} = A \cup B^c$. So the statement is false.
- (d) The statement is true, because $B \cup B^c = S$.

- E4) In case of drawing with replacement, total number of sample points in the sample space is given by
 $\zeta = \{(r_1, r_1)(r_1, r_2), (r_2, r_2)(r_1, w), (w, r_1), (r_2, w), (w, r_2), (w, w), (r_2, r_1)\}$
 No. of sample points favouring the occurrence of event E is 4, so that
 $P(E) = 4/9.$

- E5) (a) Yes, because the non-negative quantities associated with the mutually exclusive and collectively exhaustive simple events add up to 1.
 (b) $P(A) = P(0) + P(1) + P(2) = 0.58.$
 (c) $P(B) = P(4) + P(3) = 0.42$; Note also that $B = A^c$. Hence
 $P(B) = 1 - P(A).$

- E6) Let us code the books as 1, 2, 3, 4, 5 and 6. say, the damaged books coded as 1 and 2. Thus, $P(\text{he will elect the two damaged copies})$
 $\frac{2}{6} \times \frac{1}{5} = \frac{1}{15}$. Also,
 $P(\text{he will elect at least one of the two damaged copies}) = 1 - P(\text{both the books he picks up are the undamaged ones}) = 1 - \frac{4}{6} \times \frac{3}{5} = \frac{3}{5}.$

- E7) Obviously, the total number of ways in which he can answer all the 100 questions is $6 \times 6 \times 6 \times \dots \times 6 = 6^{100}$. He can choose the 20 questions to be answered correctly in ${}^{100}C_{20}$ ways. For wrong answers, he can answer each of the 80 questions in 5 ways while each of the 20 questions has one correct answer so that each such question can be answered correctly in only one way. Thus exactly 20 questions can be answered correctly and hence the remaining 80 are answered wrongly in
 ${}^{100}C_{20} 1^{20} 5^{80}$ ways. Therefore,
 $P(\text{20 questions are answered correctly}) = {}^{100}C_{20} 1^{20} 5^{80}/6^{100}$
 $= {}^{100}C_{20} (1/6)^{20} (5/6)^{80}.$

E8) There are 36 different results possible for each throw of a pair of dice. Thus, if the pair is thrown n times, the total number of possible results is 36^n . Again, there are 6 identical results on both dice each time the pair is thrown and thus these do not occur in a single throw in 30 different ways; therefore, these do not occur at all in n throws in 30^n ways. So, $P(\text{any of the six combinations } (1,1), (2,2), \dots, (6,6) \text{ appears at least once}) = 1 - (30/36)^n$.

E9) Suppose n is the required number. The probability of no head at all in these n tosses is $(1/2)^n$, so that the probability of at least one head is $1 - (1/2)^n$. We need to determine n such that $1 - (1/2)^n \geq 0.9$ i.e. $(1/2)^n \leq 0.1$ i.e. $n \ln(1/2) \leq \ln(1/10)$ i.e. $-n \ln 2 \leq -1$ i.e. $n \geq (1/\ln 2) = 3.32$; Thus, we shall take n as 4.

E10) (i) Because, the events connected with different tosses are independent

$$P(\text{sum is odd})$$

$$= P(\text{score on one die is odd and the score on the other is even})$$

$$= 2 P(\text{score on die I is odd and the score on die II is even})$$

$$= 2 P(\text{score on die I is odd}) P(\text{the score on die II is even})$$

$$= 2 (3/6)(3/6) = 3/6.$$

$$P(\text{sum of scores is 7} | \text{sum is odd})$$

$$= P(\text{sum of scores is 7} \cap \text{sum is odd}) / P(\text{sum is odd})$$

$$= P(\text{sum is 7}) / P(\text{sum is odd})$$

$$= (6/36) / (3/6) = 1/3.$$

(ii) $P(\text{the sum is greater than 6}) = 1 - P(\text{sum is 6 or less})$

$$= 1 - (15/36) = 21/36 = 7/12.$$

$$P(\text{sum of scores is 7} | \text{the sum is greater than 6})$$

$$= P(\text{sum is 7}) / P(\text{the sum is greater than 6}) = (6/36) / (21/36) = 6/21 = 2/7.$$

(iii) $P(\text{the outcome of the first die was odd}) = 3/6.$

$$P(\text{sum of scores is 7} | \text{the outcome of the first die was odd})$$

$$= P(\text{scores of (1, 6) or (3, 4) or (5, 2)}) / P(\text{the outcome of the first die was odd})$$

$$= (3/36) / (3/6) = 1/6.$$

(iv) $P(\text{the outcome of the second die was even}) = (3/6)$

$$P(\text{sum of scores is 7} | \text{the outcome of the second die was even})$$

$$= P(\text{scores of (5, 2) or (3, 4) or (1, 6)}) / P(\text{outcome of the second die was even}) = (3/36) / (3/6) = 1/6.$$

(v) $P(\text{the outcome of at least one of the dice was odd})$

$$= 1 - P(\text{outcomes of both even}) = 1 - (3 \times 3 / 6 \times 6) = \frac{3}{4}.$$

$$P(\text{sum of scores is 7} | \text{the outcome of at least one of the dice was odd})$$

$$= P(\text{scores of (1, 6) or (2, 5) or (3, 4) or (4, 3) or (5, 2) or (6, 1)}) / (3/4)$$

$$= (6/36) / (3/4) = 2/9.$$

E11) (i) $P(\text{the elder child is a boy}) = \frac{1}{2}$.

$$P(\text{both children are boys} | \text{the elder child is a boy})$$

$$= P(b, b) / \{P(b, b) + P(b, g)\} = (1/4) / (1/4 + 1/4) = \frac{1}{2}.$$

(ii) $P(\text{at least one of the children is a boy}) = 1 - P(\text{both girls}) = 3/4.$

$$P(\text{both children are boys} | \text{at least one of the children is a boy})$$

$$= P(\text{both children are boys}) / P(\text{at least one of the children is a boy})$$

$$= (1/4) / (3/4) = 1/3.$$

E12) Let K_A be the event that Kalpana gets an A grade; the events K_B , K_C , R_A , R_B and R_C are similarly defined. Given

$$(a) P(K_B) = 0.3,$$

$$(b) P(R_B) = 0.4,$$

(c) $P((K_B \cap R_B) \cup (K_B \cap R_C) \cup (K_C \cap R_B)) = 0.42$
 i.e. $P(K_B \cap R_B) + P(K_B \cap R_C) + P(K_C \cap R_B) = 0.42$
 i.e. $P(K_B)P(R_B) + P(K_B)P(R_C) + P(K_C)P(R_B) = 0.42$
 i.e. $0.12 + 3 P(R_C) + 4 P(K_C) = 0.42$
 i.e. $.3 P(R_C) + .4 P(K_C) = .30$ (17)

(d) $P(K_A \cap R_B) = .08$
 i.e. $P(K_A) P(R_B) = .08$
 i.e. $P(K_A) = .08/.4 = 0.2$
 From (17), $.3 P(R_C) = .30 - .4 P(K_C) = .30 - (.4)(1 - P(K_A) - P(K_B))$
 $= .30 - (.4)(1 - 0.2 - 0.3) = 0.10$
 so that $P(R_C) = 1/3.$

Therefore, from (a), (b), (c) & (d).

$$\begin{aligned} P(\text{at least one gets a B, but neither gets a C}) \\ = P((K_B \cap R_B) \cup (K_B \cap R_A) \cup (K_A \cap R_B)) \\ = P(K_B)P(R_B) + P(K_B)P(R_A) + P(K_A)P(R_B) \\ = (.3)(.4) + (.3)(.2) + (.2)(1/3) = 0.2467. \end{aligned}$$

E13) Let A and B be the two events. Given:

(a) $P(A \cup B) = 0.5$, (b) $P(A \cap B^c) = 3/25$, (c) $P(A^c \cap B) = 8/25$.

From (a), we have $P(A) + P(B) - P(A)P(B) = 0.5$; (18)

from (b) we have, $P(A)\{1 - P(B)\} = 3/25$, i.e. $P(A)P(B) = P(A) - 3/25$;

from (c), we have, $\{1 - P(A)\}P(B) = 8/25$, i.e. $P(A)P(B) = P(B) - 8/25$. (19)

Hence, $P(A) - 3/25 = P(B) - 8/25$, so that $P(A) = P(B) - 5/25$. Now, from (18) and (19), $P(B) - 5/25 + P(B) - \{P(B) - 8/25\} = 0.5$,

i.e. $P(B) = 0.5 - 3/25 = 0.38$ which implies $P(A) = P(B) - 5/25 = 0.38 - 0.2 = 0.18$.

Then $P(\text{none of the two events occur}) = P(A^c \cap B^c) = P(A^c)P(B^c)$

$$= \{1 - P(A)\}\{1 - P(B)\} = (.82)(.62) = 0.5084.$$

E14) Given : $0.6 = P(A \cup B) = P(A) + P(B) - P(A)P(B)$

$$= 0.4 + P(B) - 0.4 P(B)$$
, so that $P(B) = .2/6 = 1/3.$

UNIT 3 PROBABILITY DISTRIBUTIONS

Structure	Page No.
3.1 Introduction Objectives	58
3.2 Random Variable Discrete Random Variable Continuous Random Variable	59
3.3 Binomial Distribution	69
3.4 Poisson Distribution	74
3.5 Uniform Distribution	77
3.6 Normal Distribution	80
3.7 Summary	86
3.8 Solutions/Answers	86
3.9 Appendix: Statistical Tables	91

3.1 INTRODUCTION

In the first two units of this block you learnt a few methods for exploring and describing numerical data. The examples you saw there showed that numerical observations on a characteristic might need careful exploration for any useful and correct inference. In general, when you decide to collect data on a characteristic, you have a specific purpose; you want to either verify a hypothesis or want to estimate a quantity for certain purpose. For example, you may want to know if the district of Jalandhar produces more wheat in a year than the district of Faridkot in the state of Punjab. You may have a hypothesis that in general, Jalandhar district produces more wheat per year than Faridkot district. When you have such a specific question or hypothesis, you will decide to collect the appropriate information or data, in this particular instance, on the annual wheat production in the two districts of Punjab. You may find a secondary source, such as a publication from an agency or a department of the government which has already collected this information and has tabulated the data. Of course, you need to be sure about the reliability of the source from where you have obtained the required data. When you actually look at the data on the amount of annual wheat production in the two districts, for say, the past ten years, what do you think you will find? Do you think the amount produced in any district will be the same from year to year? This is very unlikely as there are a very large number of factors that influence the amount of wheat produced and these factors do not remain constant over the 10 years. Also it is very unlikely that you will find Jalandhar district producing more wheat than the district of Faridkot in each of the 10 years. In such a situation how does one compare the two districts in respect of wheat production?

Firstly, it is clear from the above discussion that a realistic model will be to assume that the annual production of wheat in a district is a random variable (result of a random experiment introduced in the previous unit). The ten year observations may be thought of as the outcome of ten repetitions of a random experiment for a district. Secondly, we need to develop measures to compare the two districts with respect to the annual production of wheat when this is thought of as outcome of a random experiment.

In certain situations or for certain problems, you may find that there is no reliable secondary source where data on the appropriate characteristic are available, and that you have to take measurements or make observations, if necessary after conducting a controlled experiment. As in practice, it is never possible to control or eliminate the influence of all the factors, this experiment may also have to be modelled as a random experiment. This need arises, for instance, when a scientist claims to have developed a variety of wheat which yields more per acre than the existing varieties and you want to verify this claim. To summarise, you notice the following:

- 1) Given a specific purpose, data are collected on an appropriate characteristic.
- 2) In most real-life problem situations, the value of the characteristic of interest may depend on a large number of factors which are not constant over space or time. It is better, then, to model the situation as a random experiment and the value of the characteristic of interest as an outcome of a random experiment.
- 3) Methods have to be developed to draw the correct conclusion or inference from the data collected on a characteristic, which is the outcome of a random experiment.
- 4) If there is no secondary source of data, you may need to either conduct a controlled experiment, or conduct a survey, to obtain data on the appropriate characteristic.

We start this unit by defining the notion of a random variable and its probability distribution in Section 3.2. The notions of discrete and continuous random variables are introduced next, followed by the notions of expectation and variance. You will see that to compare random variables or to draw inferences about them in a practical application, their probability distributions are important. Also important are certain measures of the distribution such as the mean and variance. These are discussed in Sec.3.2. In the next four sections we have discussed the most commonly used four probability distributions in detail.

Here is a list of what you should be able to do by the end of this unit.

Objectives

After reading this unit, you should be able to

- specify when a variable is a random variable and classify it as discrete or continuous.
- find the probability distribution of discrete and continuous random variables and calculate the mean and variance of these distributions and use these measures to make judgements about the real-life situation.
- describe the following distributions
 - a) binomial distribution
 - b) Poisson distribution
 - c) uniform distribution
 - d) normal distribution

and calculate the mean and variance associated with these distributions.

- distinguish between the real-life situations which can be modelled (or studied) by these distributions - when to use binomial or when to use Poisson and likewise other distributions.

3.2 RANDOM VARIABLE

Let us start with the following examples:

- 1) The number of telephone calls received by an operator in a specified interval of time.
- 2) The amount of rainfall on a day.

- 3) The outcome of throwing a die.
- 4) The scores of students in an examination
- 5) The number of misprints on a randomly chosen page of a book.
- 6) The volume of sale of a certain manufactured item in a given year.
- 7) Time to failure of a machine.

All these examples have one common feature. They describe a characteristic which associates a numerical value or number to each outcomes of a random experiment. Recall from Unit 2 that **random experiments are experiments the outcomes of which cannot be predicted in advance**. We call this characteristic a variable. This characteristic depends on the outcome of the experiment and its value cannot be predicted in advance. This variable is called a random variable. "*When a variable takes different values according to chance, (which can't be predicted before hand), it is called a random variable.*" In order to make this idea clear, consider the following example.

Example 1: Suppose we are interested in the number X of heads obtained in three tosses of a coin. Let us see what is the variable in this experiment.

For that we first find the set of possible outcomes of the experiments i.e., the sample space S . The outcomes are

$$\begin{aligned} a_1 &= \text{HHH}, & a_2 &= \text{HHT}, & a_3 &= \text{HTH}, \\ a_4 &= \text{THH}, & a_5 &= \text{TTH}, & a_6 &= \text{THT}, \\ a_7 &= \text{HTT}, & a_8 &= \text{TTT}. \end{aligned}$$

Then

$$S = \{a_1, a_2, \dots, a_7, a_8\}$$

Let us denote by $X(a_j)$ the number of heads obtained when the outcome of our experiment is a_j , where $j = 1, 2, \dots, 8$. You can easily check that

$$X(a_1) = 3, X(a_2) = X(a_3) = X(a_4) = 2, \quad (1)$$

$$X(a_5) = X(a_6) = X(a_7) = 1, X(a_8) = 0 \quad (2)$$

Then do you agree that the X maps elements of the sample space S to the values $0, 1, 2, 3$? i.e. X is a function from the space S to the set $N = \{0, 1, 2, 3\}$.

Also note that corresponding to each value, there is always some sample point or a set of sample points. For example, the set of sample points corresponding to the value '0' is the single point a_8 , whereas for 1 the set is: $\{a_5, a_6, a_7\}$. That means corresponding to each value of X , there is a subset of the sample space S .

Now you again recall from Unit 2 that an event is a subset of a sample space. Thus we note that that each value of X is associated with an event.

You can, therefore make the following identification of events corresponding to the values associated by X . Denote the event corresponding to '0' as $[X = 0]$, likewise for other values. Then

$$[X = 0] = \{a_8\}, [X = 1] = \{a_5, a_6, a_7\}$$

$$[X = 2] = \{a_2, a_3, a_4\}, [X = 3] = \{a_1\}$$

Assuming that all the sample points are equally likely, we assign probabilities of $1/8$ to each of the sample points.

From here, using the law of probability you can calculate the probabilities as follows:

$$P[X = 0] = P\{a_8\} = 1/8,$$

$$P[X = 1] = P\{a_5, a_6, a_7\} = P\{a_5\} + P\{a_6\} + P\{a_7\} = 3/8,$$

$$P[X = 2] = 3/8 \text{ and } P[X = 3] = 1/8,$$

where we read $P[X = j]$ as "probability that X equals j ."

$$P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3] = 1.$$

Summing up our discussion above, we find that a random variable X is a function defined on the set of outcomes of a random experiment and it takes on a numerical value corresponding to each outcome of the experiment. Sometimes we use the abbreviation r.v. for a random variable. Now these set of possible values is a numerical representation of the original space. The following figure may help you to see this representation.

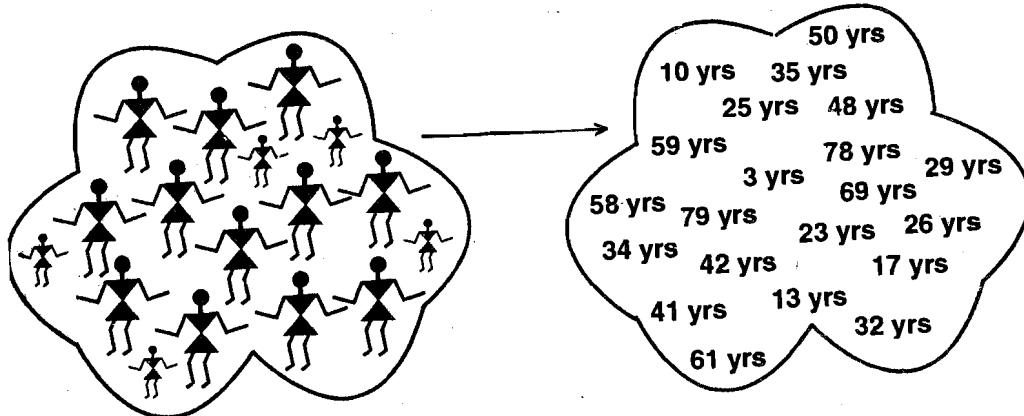


Fig. 1 Representation of a r.v.

Because of this representation, instead of working with an abstract space, we can work with a set of numbers, and this simplifies our problems considerably.

Now let us consider another example.

Example 2: You are sitting in a plane waiting for its take off. The pilot announces a delay until some incoming planes land. Suppose you want to find the following:

- i) How long will it be before take off.
- ii) How many incoming planes are there.

Let us discuss the random variables for (i) and (ii) of the above example.

We first take (i) In this case we want to find the 'duration of time' before the plane takes off. Note that the variable takes values continuously along a line as given below, say from time duration 'a' to time duration 'b'. No values in between a and b are left out. In other words there is no break in the values assumed by this random variable.



Fig.2

Now let us consider (ii). Here the random variable is the number of planes. This variable can only take the values 0 or 1 or 2 etc. as shown in Fig. 3. There is no continuity, (see Fig.3) since only non-negative integer values can be assumed.



Fig.3

The above examples show that random variables can be of different types. There are mainly two types of random variables:

- 1) **discrete random variable**
- 2) **continuous random variable**

The random variable shown in (ii) of Example 2 is discrete and that of (i) is continuous.

In the next subsection we shall discuss discrete random variables. Before that why don't you try an exercise.

- E1) Suppose you take a 50-question multiple-choice exam., guessing every answer, and are interested in the number of correct answers obtained. Then
- a) What is the random variable you will consider for this situation?
 - b) What values might this random variable have?
 - c) What would $P[X = 40]$ means?

3.2.1 Discrete Random Variables

We first formally define a discrete random variable and familiarise you with some of the properties/ aspects of a discrete r.v.

Definition 1: A random variable X is said to be discrete if the number of values that X can take is finite or countably infinite. These values may be listed as x_0, x_1, \dots , where say, $x_0 < x_1 < \dots$; these x 's are called the jump points. They need not be equidistant.

Now let us consider the events associated with the values assigned by it.

Let the events be denoted by $[X = x_i], j = 0, 1, 2, \dots$. Then, as stated earlier, we can assign probability to these events. We denote $P[X = x_j]$, the probability of the event $[X = x_j]$. For further simplification we denote the probability for each j as $P[X = x_j] = p_j, i = 0, 1, 2, 3, \dots$

From the properties of a random variable and by definition of probability it follows that

- i) $p_i \geq 0$ (for each i , i.e., each p_i is a non-negative number),
- ii) $p_0 + p_1 + \dots = 1$ (The sum of the probabilities is 1).

Now we have another definition.

Definition 2: Let $p : X \rightarrow R$ be defined as

$$p(x) = \begin{cases} p_i & \text{if } x = x_i, \quad i = 0, 1, 2, \dots \\ 0, & \text{otherwise} \end{cases}$$

Then p is called the **probability mass function**(p.m.f) of the random variable X . The collection of pairs $(x_i, p_i), i = 0, 1, 2, \dots$ is called the **probability distribution of X** .

For example, suppose X is the r.v. denoting the number of heads obtained in three tosses of a coin, then the probability mass function p is the function $p : X \rightarrow R$ such that

$$p(0) = 1/8, p(1) = p(2) = 3/8, p(3) = 1/8.$$

Note that $p(x_i) = p_i \geq 0$ for all x_i and

$$\sum p_i = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1.$$

Therefore, the probability distribution corresponding to this random variable is the set $\{(0, \frac{1}{8}), (1, \frac{3}{8}), (2, \frac{3}{8}), (3, \frac{1}{8})\}$. This is also expressed in a tabular form as follows:

Table 1: Probability distribution of number of heads in three tosses of a coin**Probability Distributions**

The number of heads (Values of the r.v. X)	Probability
0	1/8
1	3/8
2	3/8
3	1/8

Suppose you have a random variable X assuming values x_1, x_2, \dots with probabilities p_1, p_2, \dots , respectively. You may also visualise this as an illustration of a frequency distribution. In fact a frequency distribution tells you how the total probability "one" is distributed over the possible values of the random variable.

Now let us see the graphical representation of this distribution.

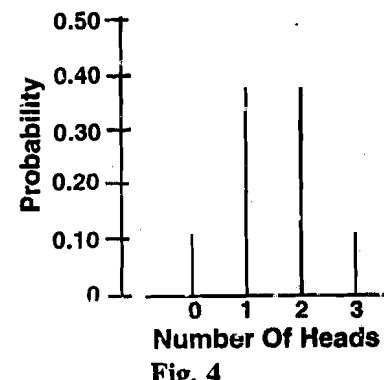
Graphically, along the horizontal axis, plot the various possible values x_i of a random variable and on each value erect a vertical line with height proportional to the corresponding probability p_i . (See Fig.4) Recall that in the bar diagram of a frequency distribution, the observed frequencies are graphed along the vertical axis; the total frequency (which is the same as the total number of repetitions of the random experiment) is thus distributed over the possible outcomes.

It is therefore clear that **associated with any random variable, a probability distribution can be defined**. Thus in the study of a random variable it is enough to know the corresponding probability distribution. This is illustrated in the following example.

Example 3: Recall the problem of Sunil, the newspaper boy, which was presented in in Section 1, Unit 2. When Sunil mentioned his dilemma about his 10 irregular customers to his sister Sunita, who is doing a course in statistics at the local college, she advised him, as a first step, to start maintaining a record for each of the 10 customers, showing for each day whether the customer has taken the newspaper from him or not. Following her advise, Sunil had at the end of two months, 61 sets of observations, each set corresponding to a day. Each set of observations was written as a sequence of two numbers, 1's and 0's, 1 at position i showing that customer i has bought the newspaper on that day and a 0 in that position showing that customer i has not bought the newspaper on that day. You will also find that the k th sequence (corresponding to the k th day) is actually an observed sample point of the sample space corresponding to the random experiment performed on day k with these 10 customers. Sunil has repeated this random experiment with his 10 customers for 61 days! When Sunil reflected over the mass of observations he has made it suddenly occurred to him that his record may be excessive for solving his problem. Do you also get the same idea as Sunil did? If not, think about it for a while now, and then read ahead.

Sunil reasoned as follows: After all, his daily gain will only depend on how many newspapers he is able to sell on a particular day, irrespective of who among the 10 buys. Therefore, it is enough for his purpose to note down the number of 1's that appear in the k th sequence corresponding to day k , to calculate his gain for that day. So why does he have to maintain a sequence? Just the total number of customers buying on a day should serve the purpose. Do you agree with Sunil? When Sunil showed his diary to Sunita and mentioned to her his new idea, she appreciated his line of thought and told him that the variable he wanted to consider, namely the total number of 1's in a sequence, would be called a random variable by a statistician as the sequences were the observed results of a random experiment. Sunil could forget about the sample space containing the sequences, provided he knew the probability distribution of the random variable chosen by him, namely the number of sales on a day to his irregular customers.

We shall consider the problem of finding the probability distribution of this random

**Fig. 4**

In the following sections you will see that there are some standard distributions and most of the real-life problems can be solved by finding the distribution corresponding to a probability model of the given situation. Let us try some exercises.

- E2) Which of the random variables given below are discrete? Give reasons for your answer.
- 1) The daily measurements of snowfall at Shimla
 - 2) The number of industrial accidents in each month.
 - 3) The number of defective goods in a shipment (lot) of goods from a manufacturer.
- E3) A box contains twice as many red marbles as green marbles. One marble is drawn at random from the box and replaced; then a second marble is drawn at random from the box. If both marbles are green, you win Rs. 50; if both are red you lose Rs. 10; and if they are of different colours, you will win or lose nothing. Then what is the probability distribution of the amount you win or lose.

Let us now return to the discussion of the three tosses of an unbiased coin. The r.v. X , denoting the number of heads obtained, has the following probability distribution:

$$p_0 = \frac{1}{8}, p_1 = \frac{3}{8} = p_2, p_3 = \frac{1}{8}$$

Suppose you want to calculate the probability of the event $\{X \leq 2\}$.

First note that the event $\{X \leq 2\}$ is the same as the event $\{X = 0\} \cup \{X = 1\} \cup \{X = 2\}$

Then, since the events are disjoint, we can write,

$$\begin{aligned} P[X \leq 2] &= P[X = 0] + P[X = 1] + P[X = 2] \\ &= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}. \end{aligned}$$

Therefore there are situations in which you are not only interested in the probability $P[X = x_i] = p_i$, but are also interested in the probability of the $X \leq x_i$, i.e., events of the type $[X \leq x_i]$ which we denote by $P[X \leq x_i]$.

As we already pointed out, the probability distribution of a random variable is analogous to a frequency distribution. It is therefore not surprising that just as we found it useful to calculate the mean of a frequency distribution, we now find it useful to calculate the mean of a probability distribution. Another term often used to denote the mean of a probability distribution is the expected value of the random variable which is defined as follows:

Definition 3: For a discrete random variable X , its **expected value** (or mean), denoted as $E(X)$, is defined as:

$$E(X) = x_0 p_0 + x_1 p_1 + \dots$$

where x_0, x_1, x_2 are the values assumed by X and p_0, p_1, p_2 are probabilities these values.

Expected value is a fundamental idea in the study of probability distributions. For many years, the concept has been put to considerable practical use in the insurance industry, and in the last twenty years, it has been widely used by many others who must make decisions under conditions of uncertainty.

We shall illustrate this idea with a real-life problem given in Example 4.

Example 4: The Director of a breast cancer screening clinic wants to know how many women will be screened on any one day. If past daily records of the clinic indicate that the number of women screened daily ranges between 100 to 115. The following table illustrates the number of times this level, between 100 to 115, has been reached during the last 100 days.

Table-2 : Number of women screened daily during 100 days

Number screened	Number of days this level was observed	Probability that the random variable will take on this value
100	1	.01
101	2	.02
102	3	.03
103	5	.05
104	6	.06
105	7	.07
106	9	.09
107	10	.10
108	12	.12
109	11	.11
110	9	.09
111	8	.08
112	6	.06
113	5	0.05
114	4	.04
115	2	.02

Total - 100

Total - 1.00

Let us see how the Director can use this past record to get information about the long-run pattern of daily patient screenings.

We hope you have understood the real-life problem. We will try to model this in statistical language. For that we shall first describe the ‘random variable’ of interest in this problem.

The random variable here is the number of patients screened, on any given day. Note that this is a discrete random variable, which can assume only non-negative integer values with positive probability. The past record of the clinic indicates that the values of this random variable range between 100 and 115 patients daily. These values are given in the 1st column of the table. The 2nd column contains the number of days each value was observed. For example the value ‘103’ occurred on 5 days. The last column gives the probability/relative frequency for which a particular value is observed. How do you calculate these probabilities? Note that the total number of days is 100 and that the value ‘100’ was observed only one day.

$$\text{Probability that the value '100' is observed} = \frac{1}{100} = .01$$

In this manner you can calculate the other probabilities. (The thumb rule is ‘divide each value in the middle column by 100’). This is how the last column is obtained. Notice that the sum of the values in the last column is one. The relative frequencies are taken as probabilities. This is the statistician’s empirical approach to assigning probabilities.

Now plot the ‘observed values’ (i.e. numbers in the 1st column) against the probabilities in a graph. Then you get a graph as given in the next page.

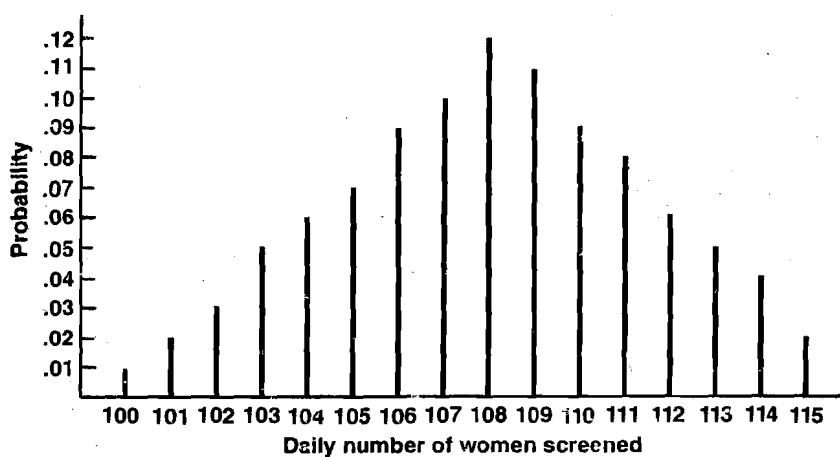


Fig.5: Probability distribution for the discrete random variable ‘number screened’

So, this is the graph of the probability distribution of the r.v.

Once we have a frequency distribution, we know that we can find its mean, variance etc. Let us calculate the mean. Here is the mean

$$\text{Mean} = \frac{\sum x_i p_i}{\sum p_i}$$

where x_i 's are the observed values in the first column and p_i 's are the frequencies

$$\text{Mean} = \frac{100 \times 1 + 101 \times 2 + \dots + 115 \times 2}{100}$$

This we can rewrite as

$$\begin{aligned}\text{Mean} &= 100 \times \frac{1}{100} + 101 \times \frac{2}{100} + \dots + 115 \times \frac{2}{100} \\ &= 100 \times .01 + 101 \times .02 + \dots + 115 \times .02 \\ &= 108.02\end{aligned}$$

So, essentially what we got is that ‘if we multiply each number in the 1st column by the corresponding number in the 3rd column’ we will get the mean of the probability distribution. This mean equals the expected value.

The above computation tells us that the expected value of the discrete random variable “number screened” is 108 women. What does this mean? It means that over a long period of time, the number of daily screenings should average about 108.02. This does not mean that 108 women will visit the clinic per day; this only says that over the **long run** and on an average 108 women can visit the clinic. This value is a long run average.

Now based on this expected value (or the mean) the director can decide on what resources/infrastructure is required to get ready for dealing with the expected number of people.

* * *

Why don't you try an exercise now.

- E4) A second-hand car dealer has sold as many as five cars in one day, and as few as one. The dealer has tabulated sales records for a large number of days and found that on 5 percent of the days no cars were sold. The dealer took 0.05 as the probability of zero sales in a day, as shown in Table 3 below. Probabilities for sales of 1,2,3,4 and 5 cars were assigned in the same manner (see Table below).

Table 3						
Number of cars sold per day	0	1	2	3	4	5
Probability	0.05	0.15	0.35	0.25	0.12	0.18

He wants to find how many cars per day will be sold on the average over a long

3.2.2 Continuous Random Variable

The variables, we discussed in the last section such as 'number of women screened', "number of heads obtained", etc. take on values $0, 1, 2, 3, \dots$. These are discrete random variables. We saw that the values of a discrete random variable are graphed as separated points and probabilities as lengths of vertical line segments (see Fig.4). We also saw that a probability distribution of such a random variable contains all possible values of the random variable, so the sum of all the probabilities must be 1.

On the contrary, a continuous random variable can take on any value in an interval.

For example, it can take all values x in the interval $0 \leq x \leq 1$ of the form 0, 0.01, 0.0002, etc. . . . , 0.98, 0.99, 1.00. As we have seen in the discrete case, we have to assign probabilities to each value of the variable. How do we do this? Since the possible values of X are uncountable, we cannot really speak of the i^{th} value of X , hence $p(x_i)$ becomes meaningless. So, what we shall do is to replace x_i by any interval of the type of (x_{i-1}, x_i) where $0 \leq x_{i-1} < x_i \leq 1$, and define probability for such intervals. To define this we consider a function defined on this interval $0 \leq x \leq 1$, which assumes non-negative values such that the total area under the graph of this function and over the interval is 1. Then we define the probability of any interval (x_{i-1}, x_i) as the area over this interval and under the graph of f . (See Fig.6(b)) At present you don't have to worry about these functions. Wherever the need arises we will specify these functions.

The above discussion may appear a little vague to you. You need not worry about this at this stage. The ideas will be clear to you once we discuss particular cases of continuous random variables and their probability distributions.

Now let us compare the graphs of probability distributions drawn for a discrete and continuous random variable.

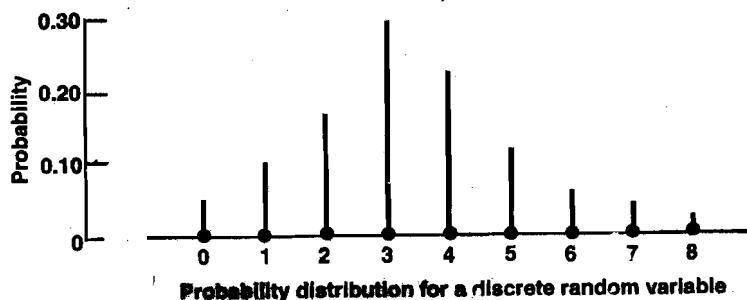


Fig.6(a)

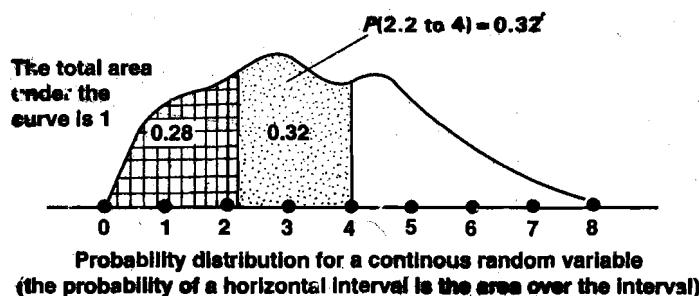


Fig.6(b)

Note that in Fig.6(b) we did not put vertical lines because a vertical distance tells us what the height of the curve is, and that is not what we want. We want areas under the portion of curves. You will see later that there are tables available for certain continuous random variables where these areas are given for different segments. For

examples in Fig. 6(b), the probability of the line segment $2.2 \leq x \leq 4$ is

$$P[2.2 \leq x \leq 4] = 0.32$$

and that of $0 \leq x \leq 2.2$ is

$$P[0 \leq x \leq 2.2] = 0.28$$

It is important to remember that a line segment has no area (zero area), because a line has no width. Thus, the vertical line segment at 4 in the distribution of Figure 6 (b) has an area of zero. This means the probability of a single value 4 is zero. In general, the **probability for an exact (single) value of a continuous random variable is zero**. Consequently, the probability of an interval is the same whether the endpoints are included or not — because the endpoints have probability zero.

Now we formalise the above discussion and make the following definition.

Definition 4: Let X be a continuous random variable which takes on values in the interval (a, b) . [i.e. all values between a and b , $a < b$]. A function $f(x)$ defined on X is called the **probability density function of X** if

- (i) $f(x)$ is nonnegative for $a \leq x \leq b$ i.e., $f(x) \geq 0$ for all x lying between a and b .
- (ii) the area under the graph and above the interval (a, b) is 1.
- (iii) For any two real numbers c and d between a and b , the probability that the random variable adopt a value between c and d is equal to the area under the graph and above the interval (c, d) i.e.

$$P[c \leq x \leq d] = \text{area under the graph over the interval } (c, d)$$

Note: Those who are familiar with the mathematical concept of integration can easily see that the above mentioned area is given by the integral of the function i.e.

$$P[c \leq x \leq d] = \int_c^d f(x)dx$$

and (ii) says that

$$\int_a^b f(x)dx = 1$$

We shall now see how we can use the graph of the distribution of a continuous random variable to study real-life problems.

Problem 1: Suppose the Director of a training programme wants to conduct a programme to upgrade the supervisory skills of production line supervisors. Because the programme is self-administered, supervisors require different numbers of hours to complete the programme. Based on a past study of participants, the following distribution (see Fig.7) showing the time spent by candidates is available which shows that average time spent is 500 hours.

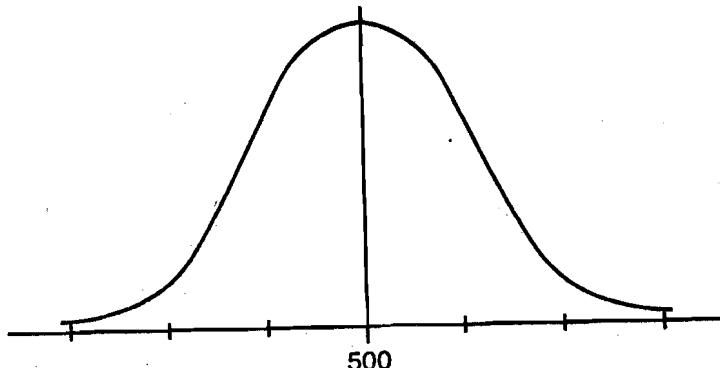


Fig.7

How can she use the graph of distribution, to find the following: What is the chance that a participant selected at random will require

i) more than 500 hours to complete the programme?

ii) less than 500 hours to complete the programme?

Solution:

- From the figure, we see that half of the area under the curve is located on either side of the mean of 500 hours. Thus, we get that the probability that the random variable will take on value higher than 500 is one half, or 0.5.
- A similar argument shows that the chance is 0.5.



Why don't you try some exercise now.

E5) Suppose X is a continuous random variable defined on [2, 4] and f is a function of [2, 4] such that

$$f(x) = \begin{cases} \frac{1}{2}, & \text{for } 2 < x < 4 \\ 0, & \text{elsewhere} \end{cases}$$

- Draw a rough sketch of $f(x)$,
- Does f define a probability density function of f? if so why?

E6) Classify the r.v.'s given at the beginning of Sec. 3.1 as discrete or continuous.

In the following sections we shall discuss some standard distribution.

3.3 BINOMIAL DISTRIBUTION

One of the important discrete random variables (or, discrete distributions) is the binomial variable. In this section we shall discuss this random variable and its probability distribution.

Many times we have to deal with experiments where there are only two possible outcomes. For example, when a coin is tossed, either head or tail comes up, seed either generates or fails to generate, a newborn is either a girl or boy.

Let us consider such an experiment. For example, consider the experiment of tossing a fair coin 3 times. This experiment has certain characteristic. First of all, it involves **repetition of three identical experiments (trials)**. Each trial has **only two possible outcomes** - a head or tail. We call outcome "head" success and outcome "tail" a failure. All trials are independent of each other. We also know that probability of getting a head in a trial and probability of getting a tail in a trial are both $\frac{1}{2}$.

$$P(H) = P(\text{success}) = \frac{1}{2}$$

and

$$P(T) = P(\text{failure}) = \frac{1}{2}$$

This shows that the **probability of a "success" and of a "failure" do not change from one trial to another**.

If X denotes the total number of heads, obtained in 3 trials, then X is a random variables which takes values from $\{0, 1, 2, 3\}$.

Suppose that p denote the probability of a success (i.e. getting a head) and q denote the probability of failure (i.e. getting a tail).

Then regarding the above experiment, we have observed the following:

- 1) It involves a repetition of n identical trials (Here $n = 3$).
- 2) The trials are independent of each other.
- 3) Each trial has two possible outcomes
- 3) The probabilities of a "success" (p) and of a "failure" (q) do not change.

If you go back for a moment to Sec. 3.1, you will see that we have already obtained the probability distribution of this in Example 3. Let us look at the probabilities once again.

$$\begin{aligned} P[X = 0] &= P[\text{getting three tails}] \\ &= P[T, T, T] = q \times q \times q \\ &= q^3 = \frac{1}{8} \end{aligned}$$

Similarly,

$$\begin{aligned} P[X = 1] &= P[[TTH], [THT], [HTT]] \\ &= P[TTH] + P[THT] + P[HTT] \\ &= q^2p + q^2p + q^2p = 3q^2p \end{aligned}$$

Similarly,

$$\begin{aligned} P[X = 2] &= P[THH] + P[THH] + P[HHT] \\ &= 3p^2q \end{aligned}$$

and

$$P[X = 3] = p^3$$

In fact the probability $P[X = r]$, $r = 0, 1, 2, 3$ gives that if we toss a coin three times, how many ways, or combinations, will yield r heads and $n - r$ tails.

Now you recall from your school mathematics that the number of combinations of n objects taken r at a time is calculated by the formula

$$c(n, r) = \frac{n!}{r!(n-r)!}$$

In the case of tossing of three coins, $n=3$. Therefore, we rewrite the probabilities as

$$\begin{aligned} P[X = 0] &= C(3, 0)p^0q^{3-0} = q^3 \\ P[X = 1] &= C(3, 1)p^1q^{3-1} = 3pq^2 \\ P[X = 2] &= C(3, 2)p^2q^{3-2} = 3p^2q \\ P[X = 3] &= C(3, 3)p^3q^0 = p^3 \end{aligned}$$

This suggests that the probability $p[X = r] = p_r$ for a given r can be calculated using the formula

$$p_r = C(n, r)p^r q^{n-r} \quad (3)$$

where r = number of successes

n = number of trials made

p = probability of success in a trial

$q = 1 - p$ = probability of failure in a trial.

Why don't you check this formula for $n=5$, i.e. tossing of a coin 5 times. For example, try this exercise.

- E7) In the experiment of tossing a coin 5 times find the probability of getting 3 heads and 2 tails. Verify that this probability is given by the formula given in Equation (3).

Let us now sum up the points we have observed in the example, above.

An experiment consisting of n trials is performed such that

- i) each trial has two possible outcomes, viz., a 'success'(p) and a "failure"(q);
- ii) the probability of success, p, is the same for any trial;
- iii) the outcomes of different trials are statistically independent (i.e. the trials are independent).

Probability Distributions

These trials are called Bernoulli trials.

The sample space of this experiment consists of elements like "SSFSSF....." of length n of 'S's and 'F's where S stands for the success and F stands for the failure.

Let X represent the number of successes (in any order whatsoever) in the set of n trials. Then X is a discrete random variable taking integral values 0, 1, ..., n. The probability of the event $P[X = r]$ is given by

Binomial Distribution

$$P[X = r] = p_r = C(n, r)p^r q^{n-r} \quad (4)$$

where r = exact number of successes

n = number of trials made

p = probability of success on a trial

$q = 1 - p$ = probability of failure on a trial and

$$C(n, r) = \frac{n!}{(n-r)!r!} \quad (\text{The earlier example illustrates how we got this formula for } p_r).$$

Such a random variable X is called a binomial random variable and its probability distribution is called binomial distribution and is given by Eqn.(2).

Problem 2: A sales representative, calls on four potential clients. The probability that she will obtain an order from each of them is $\frac{1}{2}$ and whether or not she obtains an order from one of them is statistically independent of whether or not she obtains an order from any of the others. What is the probability distribution of the number of orders she will receive?

Solution: We note that there are two mutually exclusive events (obtaining an order or no order) each time she makes a call and the probability of an order $1/2$ each time.

Also the outcomes of the calls are statistically independent. Therefore this is a situation where there are four Bernoulli trials and where the probability of a success (an order)

equals $1/2$. Substituting $n = 4$ and $p = \frac{1}{2}$ in Eqn.(4), we get that

$$P(0) = \frac{4!}{0!4!} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \frac{1}{16},$$

$$P(1) = \frac{4!}{1!3!} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \frac{1}{4},$$

$$P(2) = \frac{4!}{2!2!} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{3}{8},$$

$$P(3) = \frac{4!}{3!1!} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = \frac{1}{4},$$

$$P(4) = \frac{4!}{4!0!} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = \frac{1}{16},$$

Thus, the probability of no orders is $1/16$, of one order is $1/4$, of two order is $3/8$, of three orders is $1/4$, and of four orders is $1/16$.



James Bernoulli was a seventeenth century swiss mathematician who performed some of the early work on binomial distribution.

Problem 3: It has been claimed that in 60% of all solar heat installations, the utility bill is reduced by at least one-third. Accordingly, what are the probabilities that the utility bill will be reduced by one-third in

- i) four of five installations?
- ii) at least four of the five installations?

Solution Here the random variable follows binomial distribution with $p = 0.6$, $r = 4$ and $n=5$.

To find (i), we have to calculate $P[X = 4]$, which is given by

$$\begin{aligned} P[X = 4] &= C(5, 4)(0.6)^4(0.4) \\ &= 0.259 \end{aligned}$$

Now to find (ii), we have to find the probability that X is at least 4. This probability is the sum of the probabilities that $X = 4$ and $X = 5$ because 'at least 4 means 4 or more'. Thus we have to find $P[X = 4] + P[X = 5]$.

$$\begin{aligned} P[X = 5] &= C(5, 5)(0.6)^5 \\ &= 0.078 \end{aligned}$$

\therefore the required probability $= 0.259 + 0.078 = 0.337$.

————— X —————

Binomial distribution is very applicable in situations where we have to decide whether to accept a lot of goods (items) coming out of a manufacturing process. This decision is based on how many defective items are in the lot. Companies (or firms) will generally return the entire items if there is evidence that more than certain items is defective. To make such decision, let us see how we can make use of the binomial distribution.

An item coming out of a manufacturing process can either be defective or non defective. Consider a lot of N items produced by the manufacturing process. Let m of these be defective. Suppose a quality control inspector draws a random sample of n items from the lot, one by one, with replacement (i.e. an item drawn is put back in the lot, after noting down whether it is defective or non defective, before the next item is drawn at random). Let X be the number of defective items drawn by the inspector. Note that there are n trials and in each trial the probability that a defective item is picked remains the same, namely $\frac{m}{N}$, as the drawing is done with replacement and at random. Also note that the trials are independent. Therefore, the random variable X defined above is distributed as a Binomial (n, p) where $p = \frac{m}{N}$.

By now you must have got some idea for recognising those situations where we can apply binomial formula. If we can apply binomial distribution to study a situation, then we say that the situation can be modelled by binomial distribution.

Here are some exercises for you.

- E8) A farmer buys a quantity of cabbage seeds from a company that claims that approximately 80 % of the seeds will germinate if planted properly. If four seeds are planted, what is the probability that exactly two will germinate?
- E9) Consider again the data collected by Sunil, the newspaper boy. When Sunita, the statistics student, saw the data, she started wondering if the number of customers from among his ten irregular customers, who actually buy from him on a given day, will follow a binomial distribution? What do you think? Under what conditions will this random variable follows a binomial distribution.
- E10) Sunita was still glancing through Sunil's diary wondering to herself if she could think of the 10 customers as 'ten identical coins', when she noticed something significant. She noticed that a lot more of the sequences had a 1 in the third position than in the 8th position. Sunil remembered that customer 3 was the management trainee whom he called 'Alka Didi'. She was from a neighbouring town and was undergoing training in a software company. She was interested in news about software companies, science and environmental issues. She would often buy from Sunil but not always. Customer 8, Sunil told his sister, was a mysterious young man by name Kapil, who was rumoured to be working for a detective agency. One could rarely find him in the morning hours and if he was at

home in the morning hours, he would certainly buy from Sunil.

Probability Distributions

Given the situation above, do you still think that the number of sales on a day can be modelled as a binomially distributed random variable? Give reasons for your answer.

Once we have the probability distribution, we naturally ask what is the 'expected value'. We shall see that now.

Expected Value of a Binomial Variable

We have already seen in Sec.3.2 that for a discrete random variable X, the 'Expected Value' $E(X)$ is

$$E(X) = x_0 p_0 + x_1 p_1 + \dots$$

where x_0, x_1, \dots are the values assumed by X and p_0, p_1, \dots are the probabilities associated with these values i.e.

$$P[X = x_i] = p_i, \quad i = 0, 1, 2, \dots$$

If X is a binomial r.v., taking values $0, 1, \dots, n$, then, we know that

$$P[X = i] = C(n, i)p^i(1-p)^{n-i}$$

$$\therefore E(X) = n \times C(n, n)p^n(1-p)^0 + (n-1)C(n, n-1)p^{n-1}(1-p)^1 + \dots + 0 \times C(n, 0)p^0(1-p)^n$$

We rewrite this expression in the sum notation \sum (called sigma) as

$$\begin{aligned} E(X) &= \sum_{j=0}^n j C(n, j)p^j(1-p)^{n-j} \\ &= np \sum_{j=1}^{n-1} C(n-1, j-1) p^j(1-p)^{n-j} \end{aligned}$$

Those who are familiar with binomial expansion can recognise that the second expression on the R.H.S. is $1 - p + p^{n-1}$. Therefore we have

$$\begin{aligned} E(X) &= np[1 - p + p]^{n-1} \\ &= np. \end{aligned}$$

This means that the expected number of successes is np .

Let us do a problem.

Problem 4: An oil exploration firm plans to drill six holes. It is believed that the probability that each hole will yield oil is 0.1. Since the holes are in quite different locations, the outcome of drilling one hole is statistically independent of that of drilling any of the other holes.

(a) If the firm will be able to stay in business only if two or more holes produce oil, what is the probability of its staying in business?

(b) Give the expected value of the number of holes that result in oil.

Solution: (a) If the firm can stay in business only if two or more holes produce oil, it follows that the probability that it will stay in business equals 1 minus the probability that the number of holes resulting in oil is 0 or 1. Each hole drilled can be viewed as a Bernoulli trial where the probability of success is .1. Thus, the probability that the number of successes is 0 or 1 equals:

$$\begin{aligned} P(0 \text{ or } 1) &= P(0) + P(1) = \frac{6!}{0!6!} (.9^6) + \frac{0!6!}{1!5!} (.1)(.9^5) \\ &= .531 + .354 = .885. \end{aligned}$$

Consequently, the probability that the firm will be able to stay in business is $1 - .885 = .115$.

(b) The expected value of the number of holes yielding oil is $6 \times 0.1 = 0.6$, since $n = 6$ and $p = .1$.

A problem with the binomial distribution is that if the number trials 'n' is very large and probability 'p' is very small, computation of $P[X = r]$ is cumbersome.

The distribution which we introduce in the next section may be useful in such a situation.

3.4 POISSON DISTRIBUTION

In this section we introduce you to another discrete distribution called 'Poisson distribution'. We will familiarise you with different situations where we can apply this distribution. Let us try to understand this distribution through an example.

Poisson
A nineteenth century swiss mathematician.

Suppose it is the busy Friday noon hour at a bank, and we are interested in the number of customers who might arrive during that hour, or during a 5-minute or a 10-minute interval in that hour;

In statistical terms, we want to find the probabilities for the number of arrivals in a time interval.

As in the case of binomial, here also we make some assumptions.

- 1) The average arrival rate at any unit time remains the same over the entire noon hour.
- 2) The number of arrivals in a time interval does not depend on what happened in previous time intervals.
- 3) It is extremely unlikely that there will be more than one arrival in a very short interval of time. That means that it is impossible for more than one customer to get through the revolving entrance door in a fraction of a second.

Under these assumption we find the required probability. For this we make use of the following formula known as **Poisson formula**, given by

$$p(x) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$

where λ is the Greek letter lambda which denotes the average arrival rate per unit of time and t is the number of units of time is the number of arrivals in t units of time

Also we know that $\lambda = 72$ arrivals per hour is a constant for this situation. Since in the question λ is given in 'hour', to standardise the unit, we have to find 't' in hour.

$$\begin{aligned} \text{i.e. } 60 \text{ minutes} &= 1 \text{ hour} \\ \therefore 3 \text{ minutes} &= \frac{1}{20} \text{ hours} \\ \therefore t &= \frac{1}{20} \text{ hours} \end{aligned}$$

Then

$$p(4) = \frac{e^{-72 \times \frac{1}{20}} (72 \times \frac{1}{20})^4}{4!} = \frac{e^{-3.6} (3.6)^4}{4!}$$

To find $P(4)$, we use the Table 2, given in the Appendix. This table shows $p(x)$ for selected values of λ .

$$p(4) = 0.191$$

What does this value 0.191 specify? This tells us that if the arrivals are arrivals of customers at a bank, there is 19.1% chance that exactly four customers will arrive in the next 3 minutes.

If we vary the values of x and t , we can get different probabilities. This gives the probability distribution which is called Poisson probability distribution.

In the above discussion we saw that the Poisson formula is applicable only if certain conditions are specified. We re-state the formula now.

Poisson Formula

The Poisson Formula is given by

Poisson Distribution

$$p(x) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$

where λ is used to compute probabilities for the number of occurrences in an interval of time, if the occurrences have the following characteristic.

- 1) the average occurrence rate per unit of time is constant
- 2) occurrence in an interval is independent of what happened previously.
- 3) It rarely happens that there will be more than one occurrence in a very short time interval

A distribution having probabilities given by Poisson Formula is called Poisson distribution.

Now let us see some situations where we can apply Poisson distribution . Here is an example.

Problem 5: Calls at a telephone switch board occur at an average rate of six calls per 10 minutes. Suppose the operator leaves for a 5-minute coffee break, what is the probability that exactly two calls come in (and so go unanswered)while the operator is away?

Solution : Here you can check that the conditions 1,2 and 3 of the Poisson formula are satisfied in this case. Therefore we can use the formula. Now that here $\lambda = \frac{6}{10}$, In this case $t = 5$ so that $\lambda t = 3$. Hence the required probability $P(2)$ is given by

$$p(2) = \frac{e^3 3^2}{2!} = 3e^3 = 0.2240$$

That means there is 0.2240 chance that two calls go unanswered.

————— X —————

Here are some exercises for you.

E11) If a bank receives on an average $\lambda = 6$ bad checks per day, what is the probability that it will receive 4 bad checks on any given day.

E12) A hospital has 20 kidney dialysis machines and that the chance of any one of them malfunctioning during any day is .02. We want to find the probability that exactly 3 machines will be out of service on the same day. Then,

- i) can we use the binomial formula to find this probability? If yes, calculate the probability.

In the above exercise we have seen that the difference between the two calculations is very small.

The Poisson formula can be used to approximate the binomial probability of r successes in n binomial trials in the situations where n is large and probability of success ‘ p ’ is small.

For instance, suppose we are interested in number of road accidents in a metropolitan city or daily number of machine breakdown in a work shop etc., during a specified interval of time. Each of these subintervals is so small that at best one and no more occurrence happens within it. Thus we may look upon each subinterval as a trial. Each trial leads to a “success” if the occurrence happens during that subinterval and to a “failure” if the occurrence does not happen. Assume that the occurrences are independent of each other. Hence, the total number of occurrences can be constructed to be distributed binomially, the total number of trials being equal to the number of subintervals which we have ensured to be large; also, the length for each subinterval being small, the probability of an arrival (success) is likely to be small.

Thus we have seen that there are situations where both binomial and Poisson are applied. The rule of thumb followed by most statisticians is that if $n \geq 20$ and $p \leq 0.05$, then Poisson formula can be used to calculate binomial probability.

It is clear that the Poisson calculation is simpler than the binomial calculation. An advantage of the Poisson distribution, if it is applicable, is that it has only one parameter, λ , whereas the binomial distribution has two parameters, n and p ; consequently, Poisson probabilities can be tabulated more compactly than binomial probabilities. For example, the Poisson probability $P(3)$ is the same for $n = 200$, $p = 0.01$ as it is for $n = 100$, $p = 0.02$, and for any other pair of n and p values whose product is $\lambda = np = 2$.

By now you must have got a fairly good idea where the Poisson formula can be used.

In all the situation we have considered so far, we have calculated the probability over an interval of time. But there are situations where we need to calculate probability over a region (or space) or something else as our physical reference. In the following example we given such a situation and illustrated how to use Poisson distribution to calculate the probability.

Example 5: During second world war, a v-2 rocket hit in South London. Later a study was conducted on what are regions not affected by the rocket hit. Let us see how they used Poisson distribution for this study.

They took λ as the average number of hits per unit area (Note that earlier in the formula λ was average rate per unit time). Instead of the variable ‘ t ’ they replace the variable ‘ v ’, and x denotes the number of hits per unit area. Then they assumed that all the conditions to satisfy the Poisson formula holds in this case. With all this assumptions, they calculated the probability using the formula

$$P(x) = \frac{e^{-\lambda v} (\lambda v)^x}{x!}$$

According to the problem stated, they have to calculate the probability of ‘no hit’ per unit area. That is, the $x = 0$ and $v = 1$, so that $\lambda v = \lambda$. Now, to calculate λ , what they did was, they divided the area into 576 areas of equal size (the number 576 is chosen based on some other study and they found that they were 537 hits).

$$\therefore \text{the average number of hits per unit area } \lambda = \frac{537}{576} = 0.9323$$

Then the required Probability is

$$P[x = 0] = e^{-0.9323} = 0.3936$$

This means that if we take one region, then the probability that the region is not hit by the rocket is 0.3936. Hence, out of 576 regions, the number of regions not hit by the rocket is given by

$$576 \times 0.3936 = 226$$

* * *

Now, the actual number got from the record was that there are 229 regions not hit by the rocket. This number is quite close to 226. This shows that the values got using Poisson formula are very close to the actual values.

Thus we saw that the Poisson distribution is very effective in studying various real-life problems where the occurrence is very rare.

One of the main disadvantages of this distribution is that it is applicable only in situation where the outcomes are independent i.e. each outcome is independent of what happened previously.

In the next section we shall discuss another standard distribution.

3.5 UNIFORM DISTRIBUTION

The uniform distribution is the simplest of a few well-known continuous distributions which occur often.

As we have seen in Sec.3.2 in the continuous case we are interested in behaviour of the variable in the subintervals of the sample space, rather than at single points. If for example, the sample space is what we call the unit interval $[0, 1]$, and we set the random variable X as a value selected from this interval, then we are no longer interested in the outcome of the kind $\{X = a\}$, but rather outcome of events of the kind $\{a < x < b\}$ i.e. values lying between the two numbers a and b , where $0 \leq a \leq b \leq 1$.

Suppose X is a random variable such that if we take any subinterval of the sample space, then the probability of this interval is the same as the probability of any other subinterval of the same length. The distribution corresponding to this r.v. is called a uniform distribution. As the name suggests the probability is uniform along subintervals.

Let us see some examples of such sample spaces.

Example 6: A train is likely to arrive at a station at any time between 6.10 p.m. and 6.40 p.m. The time the train reaches, measured in minutes, after 6 p.m. is a random variable X . Here X can take any value between 10 and 40 minutes. Therefore the sample space is the interval $(10, 40)$. It is reasonable to assume that the likelihood for X taking any value between 10 and 40 is equal. So if we take subintervals of equal lengths, then the probability will be the same. The distribution corresponding to this r.v. is uniform over the interval $(10, 40)$.

* * *

Example 7: An office fire drill is scheduled for a particular day, and the fire alarm is likely to ring at any time between 9 a.m. and 5 p.m. The time the fire alarm starts, measured in minutes, after 9 a.m. is therefore a random variable which takes any value between 0 and 480 ($= 8 \text{ hours} = 8 \times 60 = 480 \text{ minutes}$) equally. The distribution corresponding to this r.v. is uniform.

* * *

Now, why don't you look for such sample spaces on your own. Try this exercise now.

- E13) Verify whether the following situations can be described by uniform distribution or not?

- a) The average life span of a life bulb produced by a manufacturing company.
 b) The number of defective items produced by an assembly process.

Next we will see how we can define (calculate) the probabilities for this distribution. As we have seen in Sec. 3.2, in the case of a continuous distribution, the probabilities are calculated using a function called ‘probability density function’ (p.d.f.). The p.d.f. for uniform distribution is given as follows.

Definition 5: The pd.f. of a random variable X which is distributed uniformly in the interval $[a,b]$, where $a < b$ is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

We can easily draw the graph of this distribution. It is given in Fig.8.

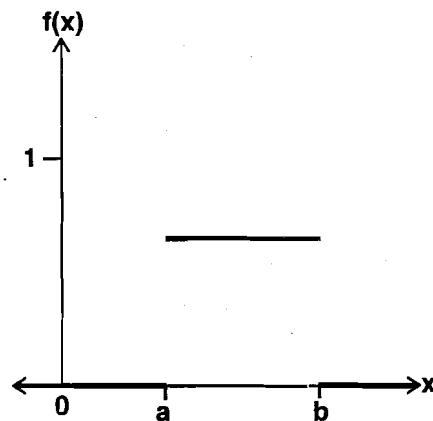


Fig. 8: Graph of P.d.f. of a uniform distribution

Now let us see how we calculate different probabilities for this distribution. As stated in Sec.3.2, for a continuous r.v., we calculate the probability of an interval rather than a point. For example, what will be $P[c < X < d]$ where $a < c < d < b$? We have seen that it is given by the area above this interval and under the graph. The area is shown in Fig.9.

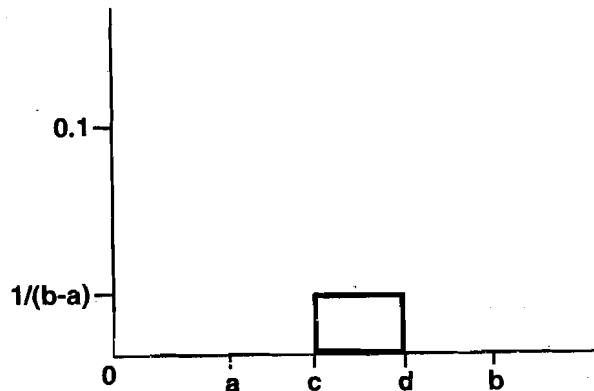


Fig.9 $P[c < X < d] = \text{Area of the rectangle shown}$

So, essentially it is the area of the rectangle with length $d - c$ and height $= \frac{1}{b-a}$ i.e.
 $P[c < X < d] = (d - c) \times \frac{1}{b-a}$

For example, if we take the situation in Example 4, let us find the probability that the

alarm sounds between 1 p.m. and 2 p.m. Here the pdf,

$$f(x) = \frac{1}{480}, 0 \leq x \leq 480$$

$$= 0, \text{ otherwise}$$

To find the required probability, you have to find time elapsed in minutes between 9 a.m. and 1 p.m. and between 9 a.m. and 2 p.m.

For 1 p.m. this is $4 \times 60 = 240$ minutes.

Similarly, for 2 p.m., it is $5 \times 60 = 300$ minutes.

Therefore you have to calculate the probability $P[240 < X < 300]$. This is given by the shaded area given below in Fig.10.

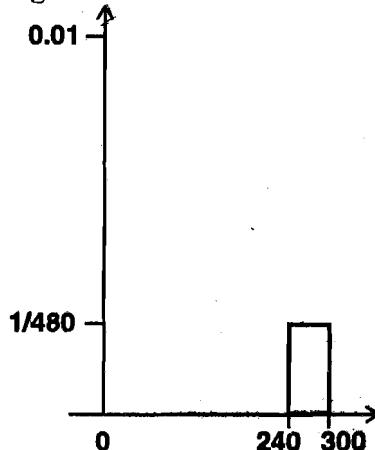


Fig.10

This area is the rectangle with base $60 (= 300 - 240)$ and height $\frac{1}{480}$.

$$P[240 < X < 300] = \frac{1}{8} = 0.125$$

That is there is 12.5% chance that the alarm sounds between 1 p.m. and 5 p.m. [Some of you may think that this fact was rather obvious from the statement of the problem itself. But we have given this situation as an illustrative example. There are situations which are complicated, where we can easily calculate the probability using this distribution.]

Next we state below the expected value of this distribution.

$$E(X) = \frac{a + b}{2}$$

You can try this exercise now.

E14) Suppose that the weight of sugar obtained by processing a tank of sugar cane juice is uniformly distributed with a mean of 10 kg. and range of 1.8 kg. Then

- i) What are the largest and smallest weights of sugar obtained from a tank of sugar can juice?
- ii) What is the probability that a tank of juice will yield sugar weighing between 9 kg. and 10.5 kg.?

E15) A train is due to arrive at 5.30 p.m. but in practise is equally likely to arrive at any time between 2 minutes early and 30 minutes late. Let the time of arrival (expressed as minutes from due time) be X. Sketch the pdf $f(x)$ of the r.v. X and shade the areas given below

- 1) The probability that the train is less than 10 minutes late.

Next we shall discuss another continuous distribution which is widely used in statistical problems.

3.6 NORMAL DISTRIBUTION

'Normal distribution' is a class of distribution which can be used to study the probability distribution occurring frequently in real-life situations, of biology, manufacturing machines, psychology etc.

A particular form of this distribution was found by seventeenth - eighteenth century mathematicians Abraham De Moivre and Pierre Laplace, while they were working on various problems in probability. They found that the distribution corresponding to certain random variables had got special property that when graphed, a bell-shaped curve is obtained and came to be called the normal pattern. The graph of the pattern became known as normal curve. Later this class of distribution was studied extensively by another mathematician Karl Friedrich Gauss and therefore this became known as 'Gaussian distribution'.

We shall now state the distribution.

Definition 6: We say that a random variable X is normally distributed with parameters μ and σ if the probability density function $f(x)$ of X is given by,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \text{ where } -\infty < x < \infty$$

where μ is a real number lying between $-\infty$ and ∞ and σ is a real number lying between 0 and ∞ .

The function $f(x)$ may look rather formidable to you at first sight. At this stage we just ask you to notice that it involves two parameters, σ and μ . Corresponding to each pair (μ, σ) , we get a distribution. Therefore there is a whole family of distributions, each one specified by a particular pair of values for σ and μ .

The most important characteristic of this distribution is that the graph of pdf, $f(x)$ for a particular value of μ and σ is bell-shaped as shown in Fig.11.

The probability density function, pdf is also symmetrical about the mean μ . The word symmetrical means that the two halves of the curve are mirror images (see Fig.11). In Fig.11 you note that if we place a mirror on the dashed vertical line (which occurs at 75 in Fig.11) then the mirror image of the portion on the left is the same as the portion on the right side.

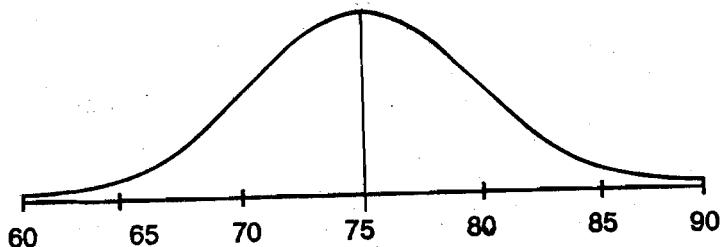


Fig.11

Both μ and σ have a 'nice' interpretation. We have already said that the pdf is symmetric about μ , so it is no surprise that μ is the **mean of the distribution**. The other constant, σ^2 dictates how spread out and flat the 'bell-shape' is and in fact σ^2 is the **variance of the normal distribution**.

As an illustration, the following figure shows that the normal pdfs for μ and σ are given as follows:

- A $\mu = 10, \sigma = 1$
- B $\mu = 10, \sigma = 2$
- C $\mu = 10, \sigma = 3$
- D $\mu = 15, \sigma = 1$

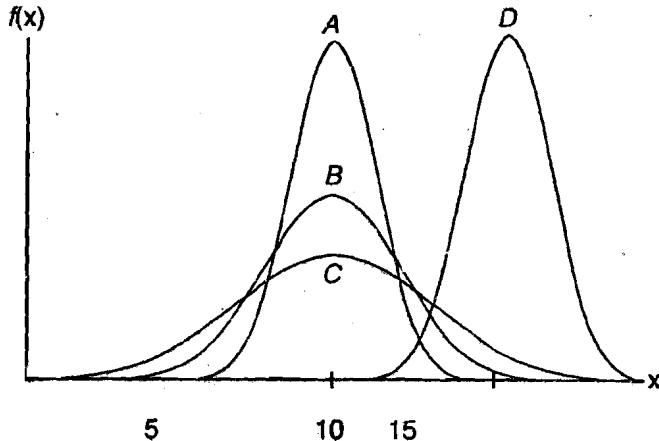


Fig.12

Pdfs A, B and C all have the mean 10 and so they are all centred at $x = 10$. Of these three curves, C has the largest variance and so is the most 'spread out'. Curve B has a smaller variance and so is less spread out, and curve A has the smallest variance and so is the most 'squeezed in'. Curves A and D have the same variance and so they have exactly the same shape, but they have different means so they are centred at $x = 10$ and $x = 15$ respectively.

Some notation

As a normal distribution is entirely specified by its parameters μ and σ we denote such distribution by $N(\mu, \sigma^2)$ where μ is the mean and σ^2 is the variance. So, for instance, the curve shown in (A) above is the pdf $N(10, 1)$ the curve in (B) is the pdf $N(10, 4)$ and so on.

The standard normal distribution

The normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$, is called the standard normal distribution. Z is the notation usually used for a random variable which has this distribution. A graph of the standard normal pdf, $p(z)$ is shown in Fig.13.

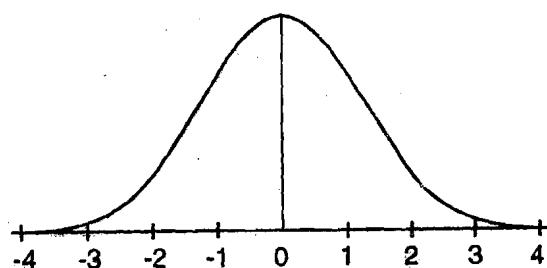


Fig.13

Notice that most of the area under the standard normal curve lies between -3 and $+3$.

Calculating Probabilities

The normal distribution is continuous and so the probability that the random variable X lies between the interval (a, b) is calculated by obtaining the area under the pdf curve between a and b.

For example, suppose an individual's IQ score, X has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$. Fig. 14 shows the areas under the pdf which correspond to $P(X < 85)$ and $P(115 < X < 120)$.

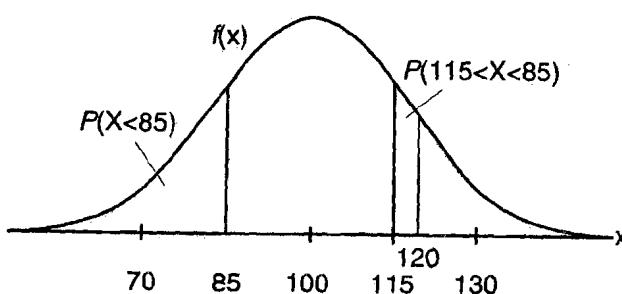


Fig.14

Unfortunately there are no 'nice' formulae for calculating such areas. But there are tables available from which we can find out the area. Statistical software are also available by which we can calculate the area.

Because the number of possible values for μ and σ is unlimited, the number of different normal distributions is unlimited. However, probabilities for every normal distribution can be obtained from a table of probability for standard normal distribution.

We shall first discuss how to use the table for calculating probabilities for a standard normal distribution. Then we shall discuss how to use this to find the probability for any normal distribution.

Using tables to calculate normal probability

We denote by $F(a) = P[Z \leq a]$, the probability that the standard normal variable Z takes values less than or equal to 'a'. The values of F for different values of a are calculated and listed in a table. One such table is given in Sec. 3.9 Appendix.

Note that the entries in the table are the values of z for $z=0.00, 0.01, 0.02, 3.49$. To find the probability that a random variable having the standard normal distribution will take on a value between a and b , we use the equation

$$P[a < z < b] = F(b) - F(a),$$

and, if either a or b is negative, we also make use of the identity

$$F(-z) = 1 - F(z).$$

In the following example we illustrate how we use the table to calculate different probabilities.

Example 8: Suppose want to find the following probability

- i) $P[0.87 < Z < 1.28]$
- ii) $P[-0.34 < Z < 0.62]$
- iii) $P[Z \geq 0.85]$
- iv) $P[Z \geq -0.65]$

We proceed as follows.

- i) We know that

$$P[0.87 < Z < 1.28] = F(1.28) - F(0.87)$$

To find $F(1.28)$, we find the row where $Z=1.2$, then move across that row to the column headed 0.08 and found the entry 0.8997. Similarly we can find that $F(0.87) = 0.8078$. Then the required Probability is

$$\begin{aligned} P[0.87 < Z < 1.28] &= 0.8997 - 0.8078 \\ &= 0.0919 \end{aligned}$$

ii) Similarly,

$$\begin{aligned} P[-0.34 < Z < 0.62] &= F(0.62) - F(-0.34) \\ &= F(0.62) - [1 - F(0.34)] \end{aligned}$$

by the identity $F(z) = 1 - F(-z)$.

$$\begin{aligned} &= 0.7324 - (1 - 0.6331) \\ &= 0.3655. \end{aligned}$$

iii) From the previous unit (Unit 2), you have already learnt that

$$P[Z \geq a] = 1 - P[Z \leq a]$$

Hence we have

$$\begin{aligned} P[Z > 0.85] &= 1 - P[Z \leq 0.85] \\ &= 1 - F(0.85) \\ &= 0.1977. \end{aligned}$$

v) As in (iii), we can write

$$\begin{aligned} P[Z > 0.65] &= 1 - P[Z \leq -0.65] \\ &= 1 - F(-0.65) \\ &= 1 - F(1 - F(0.65)) \\ &= F(0.65) \\ &= 0.7422. \end{aligned}$$

* * *

In the following exercise we ask you to find certain probabilities using the normal distribution table.

E16) If a random variable has the standard normal distribution, find the probability that it will take on a value

- i) less than 1.50
- ii) less than -1.20
- iii) greater than -1.75

E17) A filling machine is set to pour 952 ml (millimetres) of oil into bottles. The amounts of fill are normally distributed with a mean of 952 ml. and a standard deviation of 4 ml. Use the standard normal table to find the probability that a bottle contains oil between 952 and 956 ml.

Next we shall see that how to use the standard normal probability table to calculate probability of any normal distribution.

Standardising

Any normal random variable X , which has mean μ and variance σ^2 can be standardised as follows.

Take the variable X , and

- i) subtract its mean, μ and then
- ii) divide by its standard deviation, σ .

We will call the result, Z , so

$$Z = \frac{X - \mu}{\sigma}$$

For example, suppose, as earlier, that X is an individual's IQ score and that it has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$. To standardise

an individual's IQ score, X , we subtract $\mu = 100$ and divide the result by $\sigma = 15$ to give,

$$Z = \frac{X - 100}{15}$$

In this way every value of X , has a corresponding value of Z . For instance, when $X = 130$, $Z = \frac{130-100}{15} = 2$ and when $X = 90$, $Z = \frac{90-100}{15} = -0.67$.

The distribution of standardised normal random variables

The reason for standardising a normal random variable in this way is that a standardised normal random variable

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution.

That is, Z is $N(0, 1)$. So if we take any normal random variable, subtract its mean and then divide by its standard deviation, the resulting random variable will have a standard normal distribution. We are going to use this fact to calculate (non-standard) normal probabilities.

Calculating probabilities

With reference to the problem of IQ score, suppose we want to find the probability that an individual's IQ score is less than 85, i.e. $P[X < 85]$. The corresponding area under the pdf $N(100, 15^2)$ is shown in Fig.15.

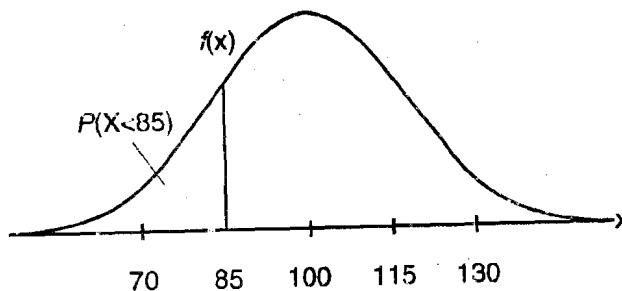


Fig.15

We cannot use normal tables directly because these give $N(0, 1)$ probabilities. Instead, we will convert the statement $X < 85$ into an equivalent statement which involves the standardised score, $Z = \frac{X-100}{15}$ because we know it has a standard normal distribution.

We start with $X = 85$. To turn X into Z we must standardise the X , but to ensure that we preserve the meaning of the statement we must treat the other side of the inequality in exactly the same way. (Otherwise we will end up calculating the probability of another statement, not $X < 85$). 'Standardising' both sides gives, $\frac{X-100}{15} < \frac{85-100}{15}$.

The left hand side is now a standard normal random variable and so we can call it Z , and we have

$$Z < \frac{85 - 100}{15}$$

which is

$$Z < -1.$$

So we have established that the statement we started with, $X < 85$ is equivalent to $Z < -1$. This means that whenever an IQ score, X , is less than 85 the corresponding standardised score, Z will be less than -1 and so the probability we are seeking, $P[X < 85]$ is the same as $P[Z < -1]$.

$P[Z < -1]$, is just a standard normal probability and so we can look it up in Table 1 in the usual way, which gives 0.1587. We get that $P[X < 85] = 0.1587$.

This process of rewriting a probability statement about X , in terms of Z , is not difficult if you are systematically writing down what you are doing at each stage. We would lay out the working we have just done for $P[X < 85]$ as follows.

X has a normal distribution with mean 100 and standard deviation 15. Let us find the probability that X is less than 85.

$$\begin{aligned} P[X < 85] &= P\left[\frac{X - 100}{15} < \frac{85 - 100}{15}\right] \\ &= P[Z < -1] = 0.1587 \end{aligned}$$

Let us do some problems now.

Problem 6: For each of these write down the equivalent standard normal probability.

- a) The number of people who visit a historic monument in a week is normally distributed with a mean of 10,500 and a standard deviation of 600. Consider the probability that fewer than 9000 people visit in a week.
- b) The number of cheques processed by a bank each day is normally distributed with a mean of 30,100 and a standard deviation of 2450. Consider the probability that the bank processes more than 32,000 cheques in a day.

Solution: Here we want to find the standard normal probability corresponding to the probability $P[X < 9000]$.

- a) We have $P[X < 9000] = P\left[\frac{X - 10500}{600} < \frac{9000 - 10500}{600}\right] = P[Z < -2.5]$.
- b) Here we want to find the standard normal probability corresponding to the probability $P[X > 32000]$.

$$P[X > 32000] = P\left[\frac{X - 30100}{2450} > \frac{32000 - 30100}{2450}\right] = P[Z > 0.78]$$

Note Probabilities like $P[a < X < b]$ can be calculated in the same way. The only difference is that when X is standardised, similar operations must be applied to both a and b . That is, $a < X < b$ becomes

$$\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}$$

which is

$$\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}$$

Problem 7: An individual's IQ score has a $N(100, 15^2)$ distribution. Find the probability that an individual's IQ score is between 91 and 121.

Solution: We require $P[91 < X < 121]$. Standardising gives

$$P\left[\frac{91 - 100}{15} < \frac{X - 100}{15} < \frac{121 - 100}{15}\right]$$

The middle term is a standardised normal random variable and so we have,

$$P\left[\frac{-9}{15} < Z < \frac{21}{15}\right] = P[-0.6 < Z < 1.4] = 0.9192 - 0.2743 = 0.6449.$$

-
- E18) A flight is due at Palam airport at 1800 hours. Its arrival time has a normal distribution with mean 1810 hours and standard deviation 10 minutes.
- What is the probability that the flight arrives before its due time?
 - Passengers must check in for a connecting flight by 1830 at the latest. What is the probability that passengers from the first flight arrive too late for the connecting flight? (Assume no travelling time from aircraft to check-in.)
- E19) The length of metallic strips produced by a machine has mean 100 cm. and variance 2.25 cm. Only strips with a weight between 98 and 103 cm. are acceptable. What proportion of strips will be acceptable? You may assume that the length of a strip has a normal distribution.
-

With this we come to an end of this unit.

Let us now summarise the points we have covered in this unit.

3.7 SUMMARY

In this unit we have covered the following points

- A random variable is a variable that takes on different numerical values according to chance outcomes
- There are two types of random variables - discrete and continuous;
- A probability distribution gives the probabilities with which the random variables take an various values in their range.
- We have discussed three standard distributions:
 - Binomial Distribution.** The probabilities of an event $P[X = r]$ in this distribution is given by
$$P[X = r] = C(n, r)P^n q^{n-r}$$
 - Poisson distribution:** The probability of an event $P[X = x]$ in this distribution is given by
$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}$$

where λ is a constant for a particular situation.

 - Uniform Distribution** The probability density function is defined by
$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{elsewhere} \end{cases}$$

The probability $P[c < X < d] = \frac{d-c}{b-a}$

 - Normal distribution.** The probability for this distribution is calculated by finding the area, under the curve of a function called probability density function defined by
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty.$$

3.8 SOLUTIONS/ANSWERS

- E1) a) If X denote the number of correct answers, then X is the random variables for this situation.

- b) X can take values $0, 1, 2, \dots$ up to 50
 c) $P[X = 40]$ means the probability that the number of correct answers is 40.
- E2) 1 and 2 are not discrete. 2 and 3 are discrete.
 (1) is not discrete because it takes values in an interval.
 (2) is discrete because the number of accidents is finite. Similarly, you argue for the situation in (3).

- E3) Let X denote the amount you win or lose. Then X takes values Rs. 50, 0 or -10 (loss in Rs. 10). The probability that both the marbles are green is $1/9$. The i.e. $P[X = 50] = 1/9$. The probability that both the marbles are red is $4/9$ i.e. $P[X = -10] = 4/9$.

The probability that the marbles are of different colour is $4/9$ i.e. $P[X = 0] = 4/9$.

Thus the probability distribution is as given in the following table.

Amount (in Rs. won (+) or lost (-))	Probability
50	$1/4$
0	$4/9$
-10	$4/9$

- E4) He has to calculate the mean. It is given by

$$\text{Mean} = \frac{0 \times 0.5 + 1 \times 0.15 + 2 \times 0.35 + 4 \times 0.12 \times 5 \times 0.18}{0.05 + 0.15 + 0.35 + 0.25 + 0.12 + 0.18}$$

$$= \frac{2.48}{1} = 2.48.$$

This means that he can expect that on an average 2 cars will be sold per day over long run (or more precisely 5 cars will be sold over (2 days)).

- E5) i)

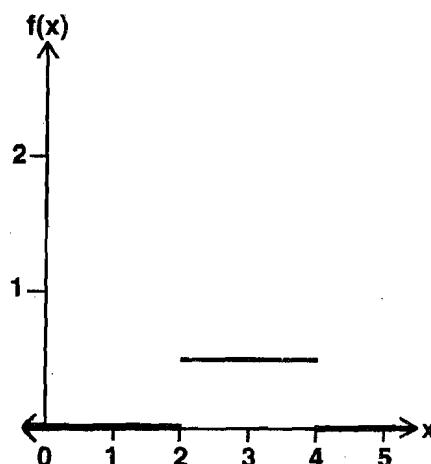


Fig.16

- ii) The area under the graph and above the interval [2, 4] is the area of the rectangle shown in Fig.16 which is given by
 $\text{Area} = 2 \times \frac{1}{2} = 1.$
 $\therefore f$ defines a probability density function of X .

- E6) 1, 3, 4, 5, 6 are discrete. 2 and 7 are continuous.

- E7) Here we have to calculate Probability of 3 heads and 2 tails. That is $P[X = 3]$. The possibility of getting 3 heads and 2 tails in five tosses of a coin is given by [HHHTT], [HHTHT], [HHTTH], [HTHHT], [HTTHH], [THTHH], [TTHHH], [THHHT], [THHTH], [THTHH], [HTHTH].

Each of these events are having probability p^3q^2 and there are 10 such events. Therefore we get

$$P[X = 3] = 10p^3q^2.$$

If we apply the formula in Equation (3) to find $P[X = 3]$, then we substitute $r = 3, n = 5$ in the formula, and we get

$$\begin{aligned} P[X = 3] &= C(5, 3)p^3q^2 \\ &= \frac{5!}{2! \times 3!} p^3 q^3 \\ &= 10p^3q^2. \end{aligned}$$

- E8) This situation follows binomial distribution with $n=4$ and $p = \frac{80}{100} = \frac{4}{5}$. The random variable X is the number of seeds that germinate. We have to calculate the probability that exactly two of the four seeds will germinate. That is $P[X = 2]$. By applying binomial formula, we get

$$\begin{aligned} P[X = 2] &= C(4, 2) \left(\frac{4}{5}\right)^2 \times \left(\frac{1}{5}\right)^2 \\ &\approx 6 \times \frac{16}{625} \\ &= 0.154 \end{aligned}$$

Therefore the required probability is 0.154.

- E9) If X_i denote the random variable that the i th customer buys the paper on a given day, then X_i 's may not be identically distributed. Therefore X_i 's may not be binomially distributed. But if the customers are having the same business activities or same kind of habits or working nature, then we can expect that X_i 's will be identically distributed. In such situation we can expect that X_i 's will follow binomial distribution.
- E10) The number of sales in a day is actually $X_1 + X_2 + \dots + X_{10}$ where each X_i is either 0 or 1 depending on whether customer i buys the paper or not on the given day. Now since customer 8 is more likely to buy on a day, than customer 3, X_3 and X_8 are not identically distributed. That is, $P[X_8 = 1] > P[X_3 = 1]$. Therefore $X_1 + X_2 + \dots + X_{10}$ cannot be thought of as binomially distributed random variable.

- E11) Since the problem deals with the receipts of bad cheques which is an event with rare occurrence over an interval of time (a day, in this case), we can apply Poisson distribution.

Since on an average 6 bad cheques are received per day,

Substituting $\lambda = 6$ and $x = 4$ in the Poisson Formula, we get

$$\begin{aligned} P[X = 4] &= \frac{6^4 e^{-6}}{4!} = \frac{1296 \times (0.0025)}{24} \\ &= 0.135. \end{aligned}$$

- E12) Note that here the experiment or trial is 'checking the machine for its functioning. There are 20 trials and each trial is identically distributed with probability 0.2.

- i) The trials are independent also. Therefore we can apply binomial formula.
We are required to calculate $P[X = 3]$. Then

$$\begin{aligned} P[X = 3] &= \frac{20!}{3! \times 17!} (0.02)^3 (0.98)^{17} \\ &= 0.0065. \end{aligned}$$

- ii) Here we have to check whether we can apply Poisson distribution.

Note that here the occurrences are 'function of the dialysis machines'. Then the average rate of machines that go out of service in a day is a constant $\lambda = 20 \times 0.02 = 0.4$.

Also note that we can make the subintervals so small that at best only one machine go out of service. Thus conditions (2) and (3) are satisfied. Therefore we

can apply Poisson Formula to calculate the required probability $P(3)$. Then

$$\begin{aligned} P(3) &= \frac{(0.4)^3 e^{-0.4}}{3!} \\ &= \frac{(0.64)(.67032)}{6} \\ &= .00715. \end{aligned}$$

Probability Distributions

E13) a) We can model it as uniform.

b) We cannot model it as uniform.

E14) i) The mean, 10 is the centre point of a line segment whose length is the range, 1.8 kg. Hence, the line segment extends $\frac{1}{2} \times (1.8) = 0.9$ kg. to the left and to the right of 10 i.e. 9.1 to 10.9 kg. Hence the smallest weight is 9.1 kg. and the largest weight is 10.9 kg.

ii) We are required to calculate $p[9 < X < 10.5]$.

$$\begin{aligned} p[9 < X < 10.5] &= (10.5 - 9) \times \frac{1}{1.8} \\ &= 0.833. \end{aligned}$$

That is the probability that the weight lies between 9.1 kg. and 10.9 kg. is 0.833.

E15) Here the pdf $f(x)$ is given by

$$\begin{cases} \frac{1}{32}, & 5.28 < x < 6 \\ 0, \text{ otherwise} & \end{cases}$$

The sketch of $f(x)$ is as given below,

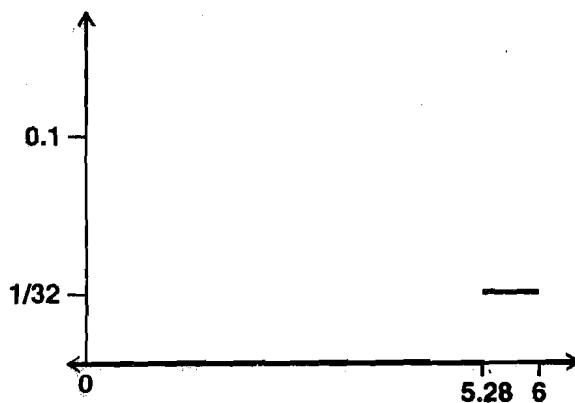


Fig.17

i) To find the required probability we note that X can take values in the interval (5.28, 5.40). Hence the required probability is

$$P[5.28 < X < 5.40] = 0.2 \times \frac{1}{32} = \frac{0.003}{8} = 0.0037$$

E16) a) 0.9332

b) 0.1151

c) 0.9599.

E17) The standard normal probability corresponding to this probability is given by

$$\begin{aligned} P[952 < X < 956] &= P\left[\frac{952 - 952}{4} < \frac{X - 952}{4} < \frac{952 - 956}{4}\right] \\ &= P[0 < Z < 1] \\ &= F(1) - F(0) \\ &= 0.8413 - 0.5 \\ &= 0.343. \end{aligned}$$

E18) Let the time of arrival in minutes past 1800hrs be X. Then X follows normal distribution $N(10, 10^2)$.

- a) The required probability is $P[X < 18]$. The standard probability corresponding to this is

$$\begin{aligned} P\left[Z < \frac{18 - 18.10}{10}\right] &= P[Z < .0] \\ &= F(0.01) \\ &= 0.5040. \end{aligned}$$

- b) The required probability is $P[X > 30]$. Then

$$\begin{aligned} P[X > 30] &= P\left[\frac{X - 10}{10} > \frac{30 - 10}{10}\right] = P[Z > 2] \\ &= 1 - P[z \leq 2] \\ &= 1 - 0.9772 = 0.0228. \end{aligned}$$

E19) We have to find the probability $P[98 < X < 103]$. The standard probability is

$$P\left[\frac{-2}{\sqrt{2.25}} < Z < \frac{3}{\sqrt{2.25}}\right].$$

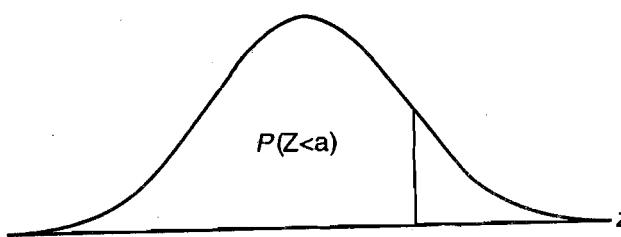
$$\begin{aligned} \text{i.e. } P\left[\frac{-2}{15} < Z < \frac{3}{15}\right] &= P[-0.13 < Z < 0.2] \\ &= F(0.2) - F(-0.13) \\ &= F(0.2) - (1 - F(0.13)) \\ &= 0.5793 - (1 - 0.5517) \\ &= 0.5793 - 0.4483 \\ &= 0.1310 \\ &= \frac{13}{100} \text{ approximately} \end{aligned}$$

So, only 13 % will be acceptable.

3.9 APPENDIX

Cumulative Standard Normal Probabilities $P[Z < a]$ where $Z \sim N(0, 1)$.

a	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000



POISSON PROBABILITY DISTRIBUTION

This table shows the value of

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

for selected values of x and for $\mu = .005$ to 8.0.

	μ									
<u>x</u>	.005	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.9950	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
1	.0050	.0099	.0192	.0291	.0384	.0476	.0565	.0653	.0738	.0823
2	.0000	.0000	.0002	.0004	.0008	.0012	.0017	.0023	.0030	.0037
3	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

	μ									
<u>x</u>	.1	.2	.3	.4	.5	.6	.7	.8	.9	.10
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0002	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	μ									
<u>x</u>	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

	μ									
<u>x</u>	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680

Continued

5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027

10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	

μ	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1734	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954

5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132

10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	

μ	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0280	.0307	.0334	.0363

Continued

x	μ									
	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002
x	μ									
	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001
x	μ									
	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014

Continued

x	μ									
	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0245	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0264
13	.0058	.0065	.0073	.0081	.0089	.0098	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001
x	μ									
	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1382	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

UNIT 4 SAMPLING DISTRIBUTIONS

Structure	Page No.
4.1 Introduction Objectives	5
4.2 Population and Samples	6
4.3 What is a Sampling Distribution	8
4.4 t-distribution	18
4.5 Chi-Square distribution	22
4.6 F-distribution	24
4.7 Summary	25
4.8 Solutions/Answers	26

4.1 INTRODUCTION

In the last block we emphasised the techniques used to describe data. To illustrate these techniques we organised the data into a frequency distribution and computed various averages and measures of dispersion. Measures such as mean and standard deviation were computed to describe the central tendency of the data and the extent of its spread. In this unit we start discussing the procedures of statistical inference which is the core of statistical analysis. We have already discussed the concepts of probability and probability distributions which lays a foundation stone for statistical inference.

As we have seen in Unit 1, in statistical inference we use the sample to make inference about the unknown population characteristic. As you know, it is a common practice for every body to draw inference about a large population by examining a sample from it. This technique is used in decision making based on statistical analysis. There if we want to make decision about a large population, may be finite or infinite, we take representative samples from the population and get results for these samples which is then used to make decisions about the population. But the results we obtain from the samples may vary from sample to sample. To make valid judgements about the population, it is necessary to study this variation. In this unit we introduce you to the concept of sampling distribution. Further we discuss how to construct a sampling distribution by selecting all samples of size, say, n from a population and how this is used to make inferences about the population. To start with we study the sampling distribution of means in Sec.4.3.

The fact that different samples are likely to give different results shows that there is a possibility of error which may affect the final decision. The sampling distribution allows us to calculate this error also.

Very often, it is not easy to determine the sampling distribution exactly. In such cases we make use of a fundamental theorem in statistics known as the Central Limit Theorem. We shall only state this theorem and discuss its utility. In Sec.4.3, we also discuss sampling distribution of proportions. You will see that this distribution is useful when we take samples from a binomial population.

In Sections 4.4, 4.5 and 4.6 we shall discuss three important sampling distributions, t , χ^2 and F .

The examples and exercises in this unit are focused on how sampling techniques can assist us in making decision about various real-life problems.

Here is a list of what you should be able to do by the end of this unit.

Objectives

After reading this unit, you should be able to

- explain the need to study the sampling distribution of a statistic,
- use sampling distribution of mean and proportion to draw inferences about the population mean and population proportion,
- calculate the error in sampling using sampling distribution,
- use central limit theorem to make inferences,
- use χ^2 , F and t-distributions to solve some problems of statistical inference.

4.2 POPULATION AND SAMPLES

The term population in statistics is applied to sets or collection of objects, actual or conceptual, and mainly to sets of numbers, measurements, or observations. For example, if we collect data on the number of television sets for each household in Mumbai, then a listing of number of television sets per household constitutes the population.

In some cases, such as concerning the number of television sets per household, the population is **finite**; in other cases, such as the determination of some characteristic of all units past, present, and future, that might be manufactured by a given process, it is convenient to think of the population as **infinite**. Similarly, we look upon the results obtained in a series of flips of a coin as a sample from the hypothetically infinite population consisting of all conceivably possible flips of the coin

In the last block we have seen that populations are often described by the distributions of the values, and it is a common practice to refer to a population in terms of this distribution. (For finite populations, we are referring here to the actual distribution of its values; for infinite populations, we are referring to the corresponding probability distribution or probability density.). For example, we may refer to a number of flips of a coin as a sample from a “binomial population” or to certain measurements as a sample from a “normal population.” Hereafter, when referring to a “population $f(x)$ ” we shall mean a population such that its elements have a frequency distribution, a probability distribution, or a density with values given by $f(x)$.

If a population is infinite it is impossible to observe all its values, and even if it is finite it may be impractical or uneconomical to observe it in its entirety. Thus, it is usually necessary to use a **sample**, a part of a population, and infer from it results pertaining to the entire population. Clearly, such results can be useful only if the sample is in some way “representative” of the population. It would be unreasonable, for instance, to expect useful generalisations about the population of 1984 family incomes in India on the basis of data pertaining to home owners only. Similarly, we can hardly expect reasonable generalisations about the performance of a tyre if it is tested only on smooth roads.

Here we notice that the elements in the population may not be always uniform in character. For example in the case of population of 1984, all of them may not be home owners. Like that all of them may not be having other characteristics like educated etc. Such population where the elements of the population are not having uniform

characteristic is called **heterogeneous population**. Otherwise the population is called **homogeneous**. So we have to take some care while taking samples from a heterogeneous population. To assure that a sample is representative of the population from which it is obtained, and to provide a framework for the application of probability theory to problems of sampling, we shall limit our discussion to what is called a random sample. For sampling from finite populations, they are defined as follows:

Definition 1 : A set of observations x_1, x_2, \dots, x_n constitutes a random sample of size n from a finite population of size N , if it is chosen so that each subset S containing n of the N elements of the population has the same probability of being selected.

Note that this definition of randomness pertains essentially to the manner in which the samples are selected. This holds also for the following definition of a random sample from a theoretical (possibly infinite) population:

Definition 2: A set of observations x_1, x_2, \dots, x_n constitutes a random sample of size n from a population with density or mass function $f(x)$ if

- 1) each x_i is a realisation of a random variable whose frequency/ density function is given by $f(x)$.
- 2) these n random variables are independent.

There are several ways of selecting a random sample. For relatively small sample, this can be done by drawing lots, or equivalently by using a table of "random numbers" specially constructed for such purposes. We shall discuss the utility of this table in and other selection procedures in detail in Unit 12, Block 4.

Before proceeding further, why don't you try some exercises now.

- E1) Suppose we want to know the average age of female students enrolled in IGNOU BDP programme. What will be the population for this study? Is this population finite or infinite?
- E2) Which of the following is an appropriate sample for studying the situation given in E1.
 - i) IGNOU female of BDP students selected from Delhi region.
 - ii) Randomly selected female students from the list of BDP students.
 - iii) Female students selected from two study centres of each regional centres.
 - iv) Randomly selected students from all students registered with IGNOU.

A population is completely determined (or characterised) by certain fixed quantities (often unknown). These fixed quantities are called **parameters** and the problem consists of "inferring" about these characteristics based on a sample data.

A function of a sample observation is called a **statistic**.

For example in the case of the problem determining the average of income daily wagers in a particular city, we find average income of a sample of workers chosen. Such measures (or quantities) calculated from a sample is called a statistic.

Try this exercise, now.

- E3) Give an example to show the difference between a parameter and a statistic.

Now, suppose we want to study the population using the "mean" as given in the earlier example. When we are referring to the mean of a population, we call it **population mean** and when we are referring to the mean of a sample, we call it **sample mean**.

Population mean is usually denoted by μ and sample mean is denoted by \bar{x}

Similarly for other measures like standard deviation, proportion etc, we have population standard deviation, sample standard deviation and population proportion, sample proportion respectively.

To understand the terms parameters and statistic in a better way, let us consider an example.

Assume that we want to draw inferences regarding the accuracy of the quantity of Milk being packed by a leading milk processing company in Western Region of India, AMUL, in 500 ml. packets.

Here the population consists of all milk packets of 500 ml. packets by AMUL company in a day. By finding out the mean of the measurements obtained from all the packets of a day's production, we get population mean. We pick up a random sample of size say n , and take the measurements. We denote the measurements as $x_1, x_2, x_3, \dots, x_n$. Note that these measurements vary with different samples. We use X_1, X_2, \dots, X_n to denote the random variables whose particular observations are x_1, x_2, \dots, x_n .

To determine the average quantity of milk being packed, we calculate the mean of the random observations, $x_1, x_2, x_3, \dots, x_n$. This is the sample mean for this sample. We denote this mean by \bar{x} . \bar{x} is a particular value of \bar{X} , where \bar{X} is

$$\bar{X} = t(X_1, X_2, \dots, X_n)$$

From the above discussion, we observe that the sample statistic, like sample mean, is a random variable. Next we shall study these random variables in detail.

4.3 WHAT IS A SAMPLING DISTRIBUTION

Let us start with situation of 'quantity of milk' discussed in the above section.

Now to study the average quantity of milk, suppose we decide that we take a sample of 10 packets without replacement and observe the quantity of milk in each packets. So here the sample size is $n=10$. The observed values are given in Column 2 Table 1 in the next page.

The mean of these 10 observations \bar{x}_1 is 496.01 for example 1. That is the average of the quantities of milk obtained from the sample is 496.01. Now, if we are using this sample to make judgements about the population, then we say that, the sample mean value is 496.01 gives an estimate of the average quantity of milk for the whole population.

Table 1

No.	Quantity of Milk in mls	
	Sample 1	Sample 2
1	502	501
2	501	493.9
3	499.5	499.6
4	501.05	490.03
5	499.05	500.09
6	497.56	500
7	501.06	500
8	459.3	499.3
9	499.6	497.5
10	500	502.09

But, if we take another sample from the population and observe the values as given in Column 3 Table 1, the mean of these values is 498.35 for sample 2. This is different

from the mean of the first sample. Now, if we are using this sample to make inference about the population, then we get a different estimate of the average quantity of milk for the whole population. Like this we can take many different samples of size 10 and each case we get sample means which may or may not be distinct. From all these values we try to estimate the mean of the whole population. In this case it is the average amount of milk in any carton.

To give you a better understanding of generalising from sample statistic to the value of the parameter, let us look at the following example in which the size of the population is very small.

Example 1: Suppose we have a population of $N=4$ incomes of four business firms and we want to find the average return of these firms. The incomes (in Lakhs) are 100, 200, 300 and 400.

We first note that in this case the (population) mean income is 250 lakhs. Now we use this situation to illustrate how sample means differ from the population mean.

Suppose we select a sample of $n = 2$ observations in order to estimate the population mean μ . Now, there are $C(4, 2) = 6$ possible samples of size 2 and we will randomly be selecting one sample from this. We shall now calculate the means of these 6 different samples. These six different samples and their means are given in the following table.

Table 2

Sample	Sample elements X_i	Sample means \bar{X}
1	100,200	150
2	100,300	200
3	100,400	250
4	200,300	250
5	200,400	300
6	300,400	350

* * *

Now, from the table above, you can find that each sample has a different mean, with the exception of third and fourth samples. Therefore four of the six samples will result in some error in the estimation process. This sampling error is the difference between the population mean μ and the sample mean we use to estimate it.

Let us now consider the possible sample means and calculate with their probability. We assume that each sample is equally likely to be chosen. Then the probability of selecting a sample is $\frac{1}{6}$

Then we list every possible sample means and their respective possibilities in a table. (See Table 3 given below).

Table 3

Sample mean \bar{X}	Number of samples yielding \bar{X}	Probability $P(\bar{X})$
150	1	1/6
200	1	1/6
250	2	2/6
300	1	1/6
350	1	1/6
		Total 1

The table obtained above is called the sampling distribution of mean.

Definition 3: A list of all possible values for a sample statistic and the probability associated with each value is called a **sampling distribution of the statistic**.

In the above example if we take the sample mean \bar{x} as the random variable, then Table 3 is nothing but the probability distribution of the means. That means, here the observed values are the means. Like any other list of numbers, these sample means have a mean. It is called '**mean of the sample means**' or the **grand mean**. The mean of the sample means is calculated in the usual fashion: the individual observations are summed, and the result is divided by the number of observation. Therefore if $\bar{\bar{X}}$ denotes this mean, then we have

$$\bar{\bar{X}} = \frac{\sum_{i=1}^k \bar{x}_i}{k}$$

where k is the number observation (samples). For the given situation of incomes, we can calculate $\bar{\bar{X}}$ from table as

$$\begin{aligned}\bar{\bar{X}} &= \frac{150 + 200 + 250 + 250 + 300 + 350}{6} \\ &= 250\end{aligned}$$

Notice that this equals the population mean $\mu = 250$. This is no coincidence. The grand mean $\bar{\bar{X}}$ will always equal the population mean. Thus we have arrived at an important observation.

If we were to take every possible sample of size n from a population and calculate each sample mean, the mean of those sample means would equal the population mean. That means $\bar{\bar{X}}$ is the population mean.

Now, why don't you calculate the mean $\bar{\bar{X}}$ for the income problem discussed, in Example 1, by taking samples of size 3 (see E5).

IV

- E4) Construct the sampling distribution for the income problem discussed in Example 1, by taking samples of size 3. Calculate the grand mean and compare it with the population mean.

After doing E4, you must have noticed that in the case of samples of size 3 also, we get that the mean of the sample means is equal to the population mean.

Note: You should not confuse n , the number of observations in a single sample, with k , the number of possible samples. In the situation of 4 incomes, the sample size is $n = 2$, while the number of possible samples is $k = 4c_2 = 6$.

Let us do a problem now.

Problem 1: A Statistics professor has given five tests. A student scored 70, 75, 65, 80 and 95 respectively in five tests. The professor decides to determine his grade by randomly selecting a sample of three test scores. Construct the sampling distribution for this process. What observations might you make?

Solution: There are ${}^5C_3 = 10$ possible samples. The samples and their means are given in Table 4.

The sampling distribution is given in Table 5. Note that the population mean is $\mu = 77$. None of these samples gives the mean as 77. Five of the ten possible samples produce values of the sample mean \bar{X} in excess of the population mean, while the other five samples underestimate it.

Table 4

Sample number	Sample elements X_i	Sample mean \bar{X}	Sample number	Sample elements X_i	Sample mean \bar{X}
1	70,75,65	70.0	6	70,80,95	81.7
2	70,75,80	75.0	7	75,65,80	73.3
3	70,75,95	80.0	8	75,65,95	78.3
4	70,65,80	71.7	9	75,80,95	83.3
5	70,65,95	76.7	10	65,80,95	80.0

Table 5

\bar{X}	P(\bar{X})
70.0	1/10
71.7	1/10
73.3	1/10
75.0	1/10
76.7	1/10
78.3	1/10
80.0	2/10
81.7	1/10
83.3	1/10
	1.00

You can try some exercises now.

E5) The ages of six executives of a company are

Name	Age
Mr. Ravi	54
Mrs. Veena	50
Mrs. Shanti	52
Mr. Suresh	48
Mr. Rajiv	50
Mr. Anil	52

- i) How many samples of size 2 are possible?
 - ii) Construct the sampling distribution of means by taking samples of size 2 and organise the data.
 - iii) Calculate the mean of the sampling distribution and compare it with the population mean.
-

As we observed in the example above , it will be interesting to note how much these sample means vary from the population mean. From your knowledge of frequency distribution (see Unit 1, Block 1), you may think that this variance can be obtained by calculating the variance of the sample means. You are right. The sample means also have a variance. It measures the dispersion of the individual observations (sample means) around their population means. Furthermore this variance is calculated like any other variance. We can obtain this by performing the following.

- 1) the amount by which each of the observations (sample means) differs from the population mean.
- 2) squaring these deviations.

3) dividing the squared deviations by the number of sample means, k,

We denote by $\sigma_{\bar{X}}$, the standard deviation of the sampling distribution of sample means. Then we have

$$\sigma_{\bar{X}}^2 = \frac{\sum(\bar{x}_i - \bar{\bar{X}})^2}{k} \quad (1)$$

Note that $\bar{\bar{X}}$ is the mean of the sampling distribution which is the same as the population mean.

Now that we have seen too many "standard deviations. One is the standard deviation of the entire population which we usually denote by σ . then the standard deviation of a single sample, and now we have the standard deviation $\sigma_{\bar{X}}$ of an entire set of sample means. Since $\sigma_{\bar{X}}$ measures the dispersion of the sample means around μ , it gives a measure for the error in sampling. Because of this, $\sigma_{\bar{X}}$ is called sampling error or standard error. We formally define this now.

Definition 3: The standard error is the standard deviation of the sample means around the population mean and is denoted by $SE(\bar{x})$.

Let us now calculate the standard errors of the sampling distribution of mean obtained in Problem 1.

Problem 2: Compute the standard error of the sampling distribution of sample scores obtained in Problem 1.

Solution: Here

$$\begin{aligned}\bar{\bar{X}} &= \frac{\sum_{i=1}^k \bar{x}_i}{k} = 77 \\ \sigma_{\bar{X}}^2 &= \frac{\sum_{i=1}^k (\bar{x}_i - \bar{\bar{X}})^2}{k} \\ \sigma_{\bar{X}}^2 &= \frac{(70 - 77)^2 + (75 - 77)^2 + (80 - 77)^2 + (71.7 - 77)^2 + \dots + (80 - 77)^2}{10} \\ &= 17.6 \\ \sigma_{\bar{X}} &= \sqrt{17.6} = 4.20\end{aligned}$$

This shows that the mean of all 10 possible sample means is 77. The square root of the average square deviation of the sample means from the population mean of 77 is 4.2.

You will study later in Unit 5 that the standard error of any statistic gives us an idea of how good a statistic is in estimating the parameters.

You may have realised that the computation of the standard deviation from the sampling distribution is a tedious process. There is an alternative method to compute standard error of the means, $SE(\bar{x})$, from a single sample if we know the population standard deviation. By this method, we have the following formula for obtaining the standard error of the mean for a finite population.

$$SE(\bar{x}) = \sqrt{\frac{(N-n)}{N-1} \frac{(\sigma^2)}{n}}$$

where N = population size = the total number of individuals, n = sample size = number of individuals selected in the random sample. σ = standard deviation of the individuals in the population.

We will not derive the formula here since the process is too technical for the scope of this course. The factor $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor. As a rule of thumb, when $\frac{n}{N}$ is less than 0.1, this correction factor can be ignored. We use the above formula for computing $SE(\bar{x})$ when $N-n$ is not very large. When N is large, relative to n , we use the formula,

$$SE(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

This formula can also be used for calculating $SE(\bar{x})$ for infinite population.

Let us see an example.

Problem 3: The U.S. Bureau of census wishes to estimate the birth rates per 1,00,000 people in the nation's largest cities. It is known that the standard deviation in the birth rates for these 100 urban centres is 12 births per 1,00,000 people. Then

- (a) calculate the variance and standard error of the sampling distribution of i) $n = 8$ cities ii) $n = 15$ cities.
- (b) compare the values obtained in both the cases.

Solution:

- (a) i) Here $N = 100$ and $n = 8$ and the population variance is 12. Therefore $\frac{n}{N}$ is less than 0.1. Then we use the formula for calculating the variance as

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{12^2}{8} = 18$$

and the standard error is $\sigma_{\bar{X}} = \sqrt{18} = 4.24$

- ii) In this case $N = 100$ and $n = 15$ and therefore $\frac{n}{N}$ is greater than 0.1. Also $\sigma = 12$. Therefore we use the formula for variance as

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \left(\frac{N-n}{N-1}\right) \frac{\sigma^2}{n} \\ &= \left(\frac{100-15}{100-1}\right) \left(\frac{12^2}{15}\right) \\ &= 8.24\end{aligned}$$

Hence, $\sigma_{\bar{X}} = \sqrt{8.24} = 2.87$

- (b) On comparing both the values we observe that, the larger sample has a smaller standard error and will tend to result in less sampling error in estimating the birth rates in the 100 cities.

————— X —————

Here is an exercise for you.

- E6) From a population of 200 observations, a sample of $n=50$ is selected. Calculate the standard error if the population standard deviation equals 22.

Thus we have seen that we can compute the standard error of the sample means if we know the population standard deviation. Here you can note one thing. Usually, we do not know σ . But it is possible to estimate σ from the sample. We will talk about this later in the next section. Using that estimate we compute the standard error by the shortcut formula given earlier.

Now we summarise our discussion about mean and standard deviation of a sampling distribution. We state the following Theorem.

Theorem 1: If a random sample of size n , say X_1, X_2, \dots, X_n is taken from a population having the mean μ and variance σ^2 , then \bar{X} = the mean of (X_1, X_2, \dots, X_n) is a random variable whose distribution has the mean μ . For samples from infinite populations the variance of the distribution is σ^2/n , for samples from finite population of size N , the variance is $\frac{\sigma^2}{n} \frac{N-n}{N-1}$

So far we have been discussing about the mean and standard deviation of the sampling distribution. Next we shall see some interesting properties regarding the shape of the sampling distribution. For that let us consider the following situation.

Suppose in an experiment 50 random samples of size $n = 10$ are taken from a population having discrete uniform distribution.

Sampling is with replacement, so that we are sampling from an infinite population. The sample means of these 50 samples are given below.

4.4	3.2	5.0	3.5	4.1	4.4	3.6	6.5	5.3	4.4
3.1	5.3	3.8	4.3	3.3	5.0	4.9	4.8	3.1	5.3
3.0	3.0	4.6	5.8	4.6	4.0	3.7	5.2	3.7	3.8
5.3	5.5	4.8	6.4	4.9	6.5	3.5	4.5	4.9	5.3
3.6	2.7	4.0	5.0	2.6	4.2	4.4	5.6	4.7	4.3

We group these means into a distribution with the classes 2.0-2.9, 3.0-3.9, ..., and 6.0-6.9, then we get the following table.

Table 6

\bar{x}	Frequency
2.0-2.9	2
3.0-3.9	14
4.0-4.9	19
5.0-5.9	12
6.0-6.9	3
	50

It is clear from this distribution as well as its histogram shown in Figure 1 that the distribution of the means is fairly **bell-shaped** even though the population itself has a discrete uniform distribution.

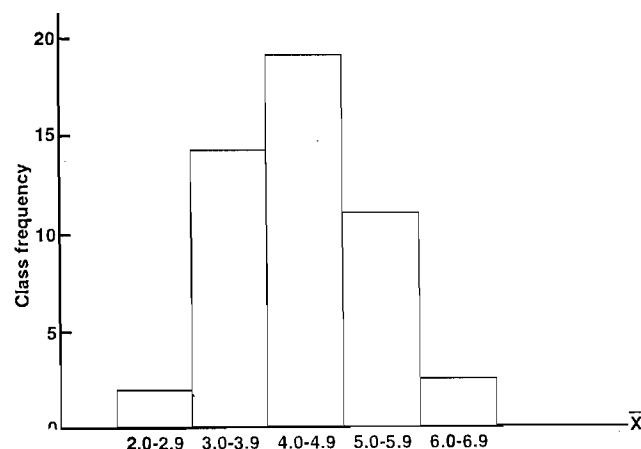


Fig. 1

This raises the question whether our result is typical of this population or any other!!

Let us consider the other situation which was discussed in the earlier section (where we want to estimate the average quantity of Milk).

Suppose in that experiment we are taking 10 random samples having size $n=10$ (In Block 4 we shall see how we can select random samples). The following table illustrates the 10 samples.

Table 7

1	2	3	4	5	6	7	8	9	10
502.0	501	493.9	497.8	502.09	502.09	502.9	493.9	459.3	497.3
501.0	493.9	499.6	499.3	499.6	501	501.9	493.09	497.56	493
499.5	499.6	502	499	490.9	500.09	501	499	493.0	499
501.05	490.03	501.3	501	496.8	500	500	499.6	499.05	499
499.05	500.09	500	502.9	500	500	500	499.9	499.5	499.3
497.56	500	499.9	501.9	503	499.3	499.3	500	499.6	500
501.06	500	493.99	500	502	499.6	499	500	500.0	500
459.3	499.3	500	499	501	493.9	499	501.3	501.0	501
499.6	497.5	502	500	493.9	497.5	493	502	501.03	501.9
500.0	502.09	499	493	503	490.03	497.8	502	501.06	502.9

If we plot these samples means, then we get the following graph.

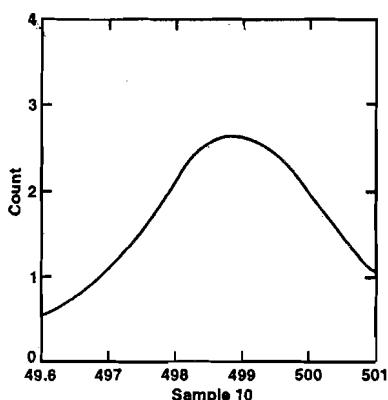


Fig.2

Here also you can observe that it is approximately bell-shaped.

Note that unlike in the earlier situation, in this case we do not know the distribution of the population.

Does the above observations give an indication that irrespective of the nature of the distribution of the population, the sampling distribution of means is approximately normal? This idea will be more clear to you if you look at the figure given at end of the unit (see Appendix 1). The figure shows graphs of the distribution function of 4 different populations see Fig. (a). The graphs in each of the figures in (b), (c) and (d) shows the sampling distribution for the sample sizes $n = 2$, $n = 5$ and $n = 30$ for the respective population. You note that the parent distribution for population 4 is normal and sampling distributions are also normal. This is not surprising. But the surprising fact is that for population 1, 2 and 3 parent distributions are not normal, still, as the sample size increases, the sampling distribution in each case approaches a normal distribution. The remarkable normality of the distribution of the sample mean is the substance of the famous central limit theorem. We state a simple version of the theorem.

Theorem 2: Central Limit Theorem : - X_1, X_2, \dots, X_n be n independent and identically distributed random variables having the same mean μ and variance σ^2 . Let $S_n = X_1 + X_2 + \dots + X_n$. Then $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ has approximately the standard normal distribution $(0, 1)$.

Now we shall illustrate this theorem in the case of sampling distribution of means (\bar{X}).

Suppose we take a sample. Let X_1, X_2, \dots, X_n , are independent random variables by the sampling scheme and also they are identically distributed. X_i 's have the same mean $\mu < \infty$ and variance $\sigma^2 < \infty$ (also called the population mean and population variance) and therefore we can apply the central limit theorem to X_1, X_2, \dots, X_n and

conclude that $\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma}$ is distributed approximately as standard normal. But note that $\frac{X_1 + \dots + X_n}{n}$ is nothing but the sample mean \bar{X} , therefore according to the central limit theorem we have $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows standard normal distribution $N(0, 1)$.

So far we have been discussing the sample means and their distribution. Our interest has been in some variable which might be measured and averaged. However there are many instances where we may need some other statistic, other than mean, to make inference about the population. In the next section we shall discuss such a statistic known as 'proportion' and discuss its sampling distribution.

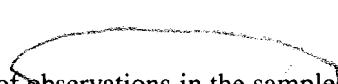
Sampling Distribution of Proportion

There are many instances where we are interested in determining the number or proportion of observations falling in a particular category. For example, a doctor wants to know how many of his patients can survive after administering a particular drug. A politician may want to know how many voters will not vote for him. In all these situations there are two possible outcomes of the observation - whether an observation falls in a particular category or not. You might recall that such situations are related to the problems dealing with binomial distribution. The politician may not be interested in the actual number of people who are going to vote for him, she is interested on what percentage of the people will do so. In such a situation we deal with sample proportions instead of sample means. The population parameter in this case is the proportion. We denote this by π . In general, for a finite population, we define the population proportion π as

$$\pi = \frac{k}{n}$$

where k is the number of observations that fall in a particular category and n is the total number of observation. When the population is very large, we may take samples to study the population and for each sample we calculate the sample proportion, p , as

$$p = \frac{s}{n}$$

 where s denotes the number of observations in the sample which meet the particular characteristic, under study and n is the sample size.

For example, assume that a politician surveys 1,000 voters and finds that only 350 are going to vote for him. Then

$$p = \frac{350}{1000} = 0.35$$

You note that, a different sample of $n = 1,000$ voters may yield a different p . If we calculate the possible sample proportions then a **list** of these observations is called the **sampling distribution of proportions**.

Let us consider an example. As, in the case of sample means, we are considering a simple situation where the population size is very small.

Example 2: In our discussion about the quantification of milk packets, let us put the condition that if the quantity in any packet measured is less than 495ml, it will be rejected by the consumer, hence call them as defective and otherwise they are non-defective.

For the data given in Table 7 we consider the first five columns (five cartons) as population then the population has 50 milk packets.

From the population we get the proportion of defective as

$$P(\text{Defective}) = \pi = \frac{7}{50} = 0.14$$

and the proportion of non-defective is

$$P(\text{non-defective}) = 1 - \pi = 0.86$$

Let us try to use the sampling method to find this proportion. Assume that random samples of size 2 out of 5 (that is two cartons out of five cartons) have been selected and the possible samples and their defective proportions are as given below

Table 8

Sample	Sample proportions (p_i)
C_1, C_2	2/20
C_1, C_3	3/20
C_1, C_4	2/20
C_1, C_5	3/20
C_2, C_3	3/20
C_2, C_4	2/20
C_2, C_5	3/20
C_3, C_4	3/20
C_3, C_5	4/20
C_4, C_5	3/20

Let us calculate the mean of the sample proportion \bar{p}

$$\bar{p} = \frac{\sum p_i}{k}$$

$$= \frac{28}{20 \times 10}$$

$$= .14$$

This is same as the population proportion.

* * *

From the earlier example we conclude the following fact:

The mean of the sampling distribution of proportions is equal to the population proportion

Now the standard error of the sample proportions, by definition, is the standard deviation of this sampling distribution. As we have mentioned earlier, in the case of mean, the computation of standard deviation from the table is a tedious process. So we make use of a short cut formula by which we can compute the **standard error** which we denote by **SE(p)**. If we know the population proportion, population size and sample size, the formula is

$$SE(p) = \sqrt{\frac{(N - n)\pi(1 - \pi)}{(N - 1).n}}$$

where π is the population proportion and N and n are the population size and sample size, respectively. If the population is infinite or if the size of the population, relative to the sample size is extremely large, then

$$SE(p) = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Let us now calculate the standard error for the situation in Example 3. In this case $\pi = 0.4$, $N = 5$ and $n=2$. Therefore,

$$\begin{aligned} \text{SE}(p) &= \sqrt{\frac{(0.14)(0.86)}{2}} \sqrt{\frac{5-2}{5-1}} \\ &= 0.2454 \times 0.866 \\ &= 0.2125. \end{aligned}$$

Here is an exercise for you.

- E7) a) In a manufacturing process of electric bulbs, company gives an output of 250 electric bulbs per day. To test the defectiveness of the manufactured bulbs, a random sample of 50 has selected in a particular day's production. Inspection of the samples showed 4 as defective and the rest as non-defective. Calculate the standard error of sample proportion of defective.
- b) If the process is considered to be a continuous one (observed for many days) and a random sample of 60 has been selected from the produced electric bulbs and observed that 7 are defective. What will be the standard error of sample proportion of defective.

So far we have discussed sampling distributions of two statistics \bar{x} and \bar{p} . We also emphasised the role of Central limit theorem and the idea that distributions of \bar{x} and p , will be approximately normal even when the data in the parent population are not distributed normally, provided the sample sizes are large.

But there are many situations where we may have to deal with small samples, may be due to limited availability of items or by other factors such as time of cost. We may also have to deal with certain statistics other than \bar{x} and p . For example in some situations we may be interested in the distribution of \bar{x}/s , where s is the sample standard deviation. In some other situations, we may be interested in the distribution of $\sum_{i=1}^n (x_i - \bar{x})^2$. The important exact sampling distributions are chi-square, student t- and F-distributions. These distributions are widely used in statistical analysis. In the following sections we shall discuss these distributions.

4.4 t-DISTRIBUTION

William G. Gosset, a brew master for Guiness Brewarier, developed a family of distributions each of which corresponds to a parameter ν , (called nu-a positive integer). The density function of a random variable whose distribution belongs to this family is given by

$$f_\nu(y) = \frac{1}{\Gamma(\pi\nu)} \frac{\Gamma(\nu + 1/2)}{\Gamma(\nu/2)} \left(1 + \frac{y^2}{\nu}\right)^{-\nu/2}$$

where $\Gamma(x)$ for any $x > 0$ is called the gamma function. The values of $\Gamma(x)$ for different values of $x > 0$ are tabulated. The three members of the family of this distribution for $\nu = 9$, $\nu = 14$ and $\nu > 30$ are shown in the accompanying figure.(See Fig. 3 in the next page.)

Now we shall illustrate the importance of this distribution as distribution in the case of sampling distribution.

Suppose we take a sample from a population having normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. If X_1, X_2, \dots, X_n are n observations in the sample, then X_1, \dots, X_n are independently and identically distributed(i.i.d) random variables. and

we have $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ has t-distribution with parameter $\nu = n - 1$, where s is the sample variance. The parameter ν is called the number of degrees of freedom (in short d.f) of the distribution.

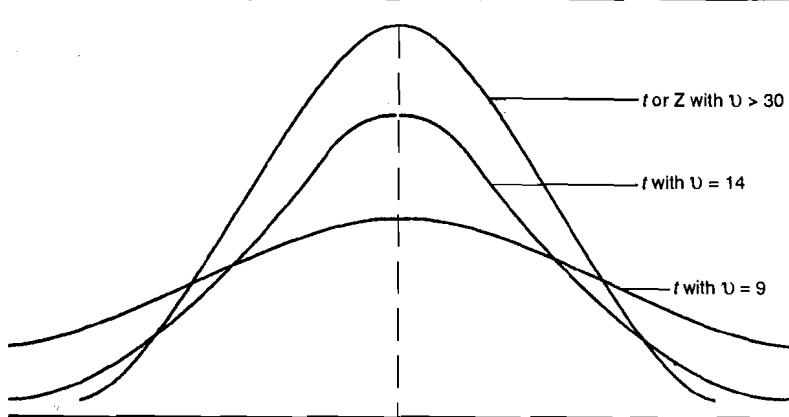


Fig. 3 t-distributions

Now that, you have been introduced to the distribution, we shall see how we use this distribution to make judgements about the population.

As we have seen in the case of normal distribution (see Unit 3, Block 1), there are tables available using which we can calculate probability for this distribution. One such table is given in the appendix (see Table 2 in appendix of this block). The table contains selected values of t_α for various values of ν , where t_α is such that the area under the t-distribution to its right is equal to α . In this table the left-hand column contains values of ν , the column headings are areas α in the right-hand tail of the t-distribution, and the entries are the values of t_α (See Fig. 4 below)

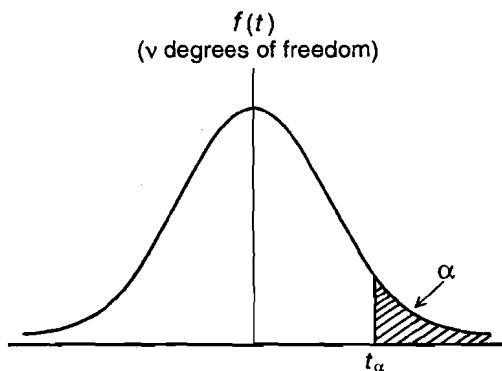


Fig. 4: Tabulated values of t

It is not necessary to tabulate the values of t_α for $\alpha > 0.50$, as it follows from the symmetry of the distribution that $t_{1-\alpha} = -t_\alpha$; thus, the value of t that corresponds to a left-hand tail area of α is t_α . You note that the bottom row of the table the entries are the same as the values of the standard normal variate 2. For example $t_{0.025} = 1.96 = z_{0.025}$ for large values of ν i.e. $\nu \geq 30$.

Remark: Please note that the tables for the distribution can vary from book to book because of the definitions of parameters involved. Therefore, for the purposes of this course, please only refer to the table given at the end of the block.

Example 3: The graph of a t-distribution with 9 degrees of freedom is given below. Let us find the values of t_1 , for which

- i) the shaded area of the right = 0.05

- ii) the total shaded area = 0.05

iii) the total unshaded area = 0.99

iv) the shaded area on the left = 0.01

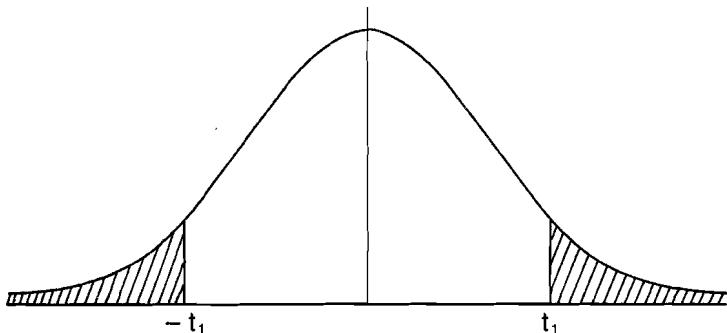


Fig. 5

Let us try (i) to (iv) one by one.

- i) If the shaded area on the right is 0.05, comparing the figure above with Fig.4, we get that $\alpha = 0.05$. It is already given that $\nu = 9$. Referring to the table given in the Appendix, we proceed downward under column headed ν until entry 9 is reached. Then proceed right to the column headed 0.05. and get the required value of t as 1.833.
- ii) Given that the total shaded area is 0.05. Then by symmetry, the shaded area on the right is $\frac{0.05}{2} = 0.025$ i.e. $\alpha = 0.025$. Also $\nu = 9$. Therefore from the table we get the required value as 2.262.
- iii) If the total unshaded area is 0.99, then the total shaded area is $1 - 0.99 = 0.01$. Therefore the shaded area on the right of t_1 is $\frac{0.01}{2} = 0.005$. Then we get the required value from the table as 3.250.
- iv) If the shaded area on the left is 0.01, then by symmetry the shaded area on the right is also 0.01. Then from the table we get the value as 2.821

* * *

Problem 4: For the given sample sizes and the t -values, find the corresponding probability α

- i) $n = 26, t = 2.485$
- ii) $n = 14, t = 1.771$

Solution:

- i) Referring to the t -table given in the Appendix, we find the row corresponding to $\nu = 25$ and look for the entry $t = 2.485$. Then we observe that this t -value lies in the column headed $t_{0.01}$. This gives that the corresponding α -value is 0.01.
- ii) Similarly you can see that α -value corresponding to $\nu = 13$ and $t = 1.771$ is 0.05.

————— X —————

Example 4: A manufacturer of fuses claims that with a 20 % overload, his fuses will blow in 12.40 minutes on the average. To test this claim, a sample of 20 of the fuses was subjected to a 20 % overload, and the times it took them to blow had a mean of 10.63 minutes and a standard deviation of 2.48 minutes. if it can be assumed that the data constitute a random sample from a normal population, do they tend to support or refute the manufacturer's claim? Let us see how we can make use of the t -distribution to find out an answer for this.

We first note that the sample is taken from a normal population and the size is small, $n = 20$. Then as we have stated in the discussion above, the random variable $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ follows t-distribution with $n - 1$ degrees of freedom.

Here $\bar{x} = 10.63$, $\mu = 12.40$ and $s = 2.48$ and $n = 20$. Therefore,

$$t = \frac{10.63 - 12.40}{2.48/\sqrt{20}} = -3.19$$

Now we compare this t-value with the values of t given in the table for the parameter, $\nu = 20 - 1 = 19$. We look for the t-value which is nearest to 3.19 and less than 3.19. Then we find that the required t-value is 2.861. Also from the table, we find that the α -value corresponding to 2.861 is 0.005. That means the probability that t will exceed 2.861 is 0.005. Now we note that -3.19 is less than -2.861 therefore the α -value corresponding to -3.19 will be less than 0.005, which is very small (see Fig.6).

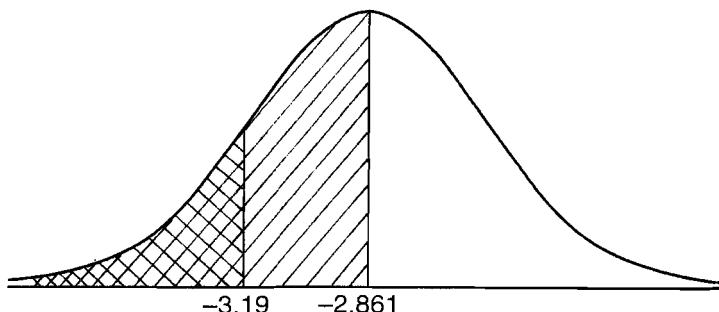


Fig. 6

Hence, we can conclude that the data tend to refute the manufacturer's claim.

* * *

Try these exercises now.

E8) Find the values of t for given values of ν and α

- i) $\alpha = 0.01$ and $\nu = 19$
- ii) $\alpha = 0.05$ and $\nu = 6$
- iii) $\alpha = 0.025$ and $\nu = 23$
- iv) $\alpha = 0.1$ and $\nu = 10$
- v) $\alpha = 0.005$ and $\nu = 29$

E9) A process for making certain bearings is under control if the diameters of the bearings have a mean of 0.5 cm. What can we say about this process if a sample of 10 of these bearings has a mean diameter of 0.5060 cm. and a standard deviation of 0.0040 cm. Assume that the data constitute a random sample from a normal population.

E10) Find the values of t for which the area of the right-hand tail of the t-distribution is 0.05 if the number of degrees of freedom is equal to

- (a) 16, (b) 27, (c) 200

Note that for applying t-distribution we made the assumption that the samples should come from a normal population and population standard deviation is not known.

Studies have been shown later that the first assumption can be relaxed. It has been shown that the distribution of random variable with the values $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ is fairly close

to a t-distribution even for samples from certain non-normal population. In practice, it is necessary to make sure primarily that the population from which we are sampling is approximately bell-shaped and too skewed.

Next we shall consider another exact sampling distribution.

4.5 CHI-SQUARE DISTRIBUTION

Another exact sampling distribution which is very useful in statistical problems is the chi-square distribution. In this section we shall introduce you to this distribution.

We say that a random variable Y has χ^2 (called chi-square) distribution with ν degrees of freedom if the density function $f_Y(y)$ of Y is given by

$$f_Y(y) = \frac{y^{(\nu/2-1)} e^{-y}}{\Gamma(\nu/2)}, \text{ if } y > 0 \\ = 0, \text{ if } y \leq 0$$

where $\Gamma(x)$ is the gamma function.

Note that this family of distributions is parametrised by ν , called the degrees of freedom. The graph of the density function for some members of this family ($\nu = 1, 3, 8, 10$) are shown in Fig.7 below

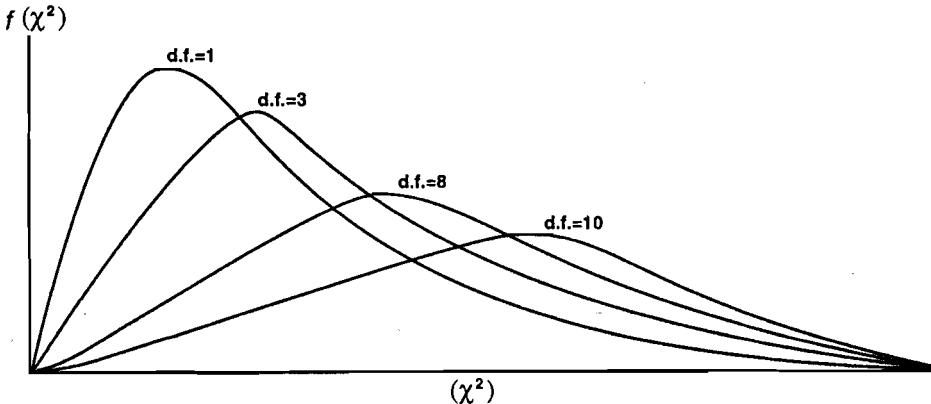


Fig.7 Various Chi-square distributions

As in the case of normal and t-distributions, a table containing selected values of χ_{α}^2 for various values of ν , again called the number of degrees of freedom, is given in the Appendix (see Table 3). χ_{α}^2 is the value such that the area under the chi-square distribution to its right is equal to α . (i.e. the probability that any value is greater than equal to or χ_{α}^2 is α). In this table the left-hand column contains values of ν , the column headings are areas α in the right-hand tail of the chi-square distribution, and the entries are the values of χ_{α}^2 . (see Fig.8 below).

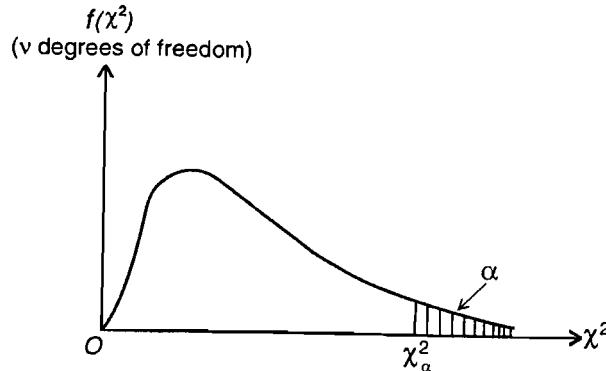


Fig.8

Note that unlike t-distribution, chi-square distribution is not symmetrical. Therefore, it is necessary to tabulate values of χ_{α}^2 for $\alpha > 0.50$ also.

Let us see some examples

Example 5: Let us find the values of χ^2_α of the χ^2 -distribution with 5 degrees of freedom for $\alpha = 0.05$ and 0.01 .

Let us first consider $\alpha = 0.05$. To find the values of χ^2_α for 5 degrees of freedom and $\alpha = 0.05$, we refer to Table 3 in the Appendix. We proceed downward under column headed ν until entry 5 is reached, then proceed right to column headed $\alpha = 0.05$ which gives the required value as $\chi^2_{0.05} = 11.070$.

Similarly for $\alpha = 0.01$, we get the required value as $\chi^2_{0.01} = 15.056$.

* * *

Now we shall illustrate the importance of χ^2 -distribution in the case of sampling distribution.

Suppose we take a random sample of size n when n is small ($n \leq 30$) from a normal population having the variance σ^2 and if s^2 is the variance of the random sample, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

is a value of a random variable having χ^2 -distribution with $\nu = n - 1$.

Let us see how we use this result for solving the following problem.

Problem 5: An optical firm purchases glass to be ground into lenses, and it knows from past experience that the variance of the refractive index of this kind of glass is $1.26 \cdot 10^{-4}$. As it is important that the various pieces of glass have nearly the same index of refraction, the firm rejects such a shipment if the sample variance of 20 pieces selected at random exceeds $2.00 \cdot 10^{-4}$. Assuming that the sample values may be looked upon as a random sample from a normal population, what is the probability that a shipment will be rejected even though $\sigma^2 = 1.26 \cdot 10^{-4}$?

Solution: We first note that the sample is taken from normal population and the size is n small, $n = 20$. Then as (we stated in the discussion above), the random variable

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \text{ follows } \chi^2\text{-distribution with } (n-1) \text{ degrees of freedom.}$$

Here $\sigma^2 = 1.26 \times 10^{-4}$, $s^2 = 2 \times 10^{-4}$ and $n = 20$. Therefore

$$\chi^2 = \frac{(10-1) \times 2 \times 10^{-4}}{1.26 \times 10^{-4}} = 30.2.$$

Now we compare this χ^2 -value with the values of χ^2 given in the table for the parameter $\nu = 20 - 1 = 19$. Then we find that the χ^2 -value 30.1 is close to the calculated value 30.2 and less than the calculated value. The corresponding α -value is 0.05. That means that the probability that χ^2 -value exceeds 30.1 is 0.05. Therefore we can conclude that the probability that a good shipment will erroneously be rejected is less than 0.05.

————— × —————

Here is an exercise for you.

E11) Find the values of χ^2 for which the area of the right-hand tail of the χ^2 -distribution is 0.5, if the number of degrees of freedom ν is equal to (a) 15, (b) 21.

E12) A random sample of 10 observations is taken from a normal population having the variance $\sigma^2 = 42.5$. Find the approximate probability of obtaining a sample standard deviation between 3.14 and 8.94.

4.6 F-DISTRIBUTION

R.A. Fisher, a British statistician, developed this distribution in the early 1920's. It is defined as follows.

Definition : If a random variable Y_1 has a χ^2 -distribution with ν_1 degrees of freedom, and if a random variable Y_2 has a χ^2 -distribution with ν_2 degrees of freedom and if Y_1 and Y_2 are independent, then $Y_1/\nu_1 \div Y_2/\nu_2$ has an F-distribution with ν_1 and ν_2 degrees of freedom.

In other words, the ratio of two independent χ^2 random variables, each divided by its number of degrees of freedom, is an F random variable.

Like the t and χ^2 distributions, the F distribution is in reality a family of probability distributions, each corresponding to certain numbers of degrees of freedom. But unlike the t and χ^2 distributions, the F distribution has two numbers of degrees of freedom, not one. Figure shows the F distribution with 2 and 9 degrees of freedom. As you can see, the F distribution is skewed to the right. However, as both numbers of degrees of freedom become very large, the F distribution tends toward normality. As in the case of the χ^2 distribution, the probability that an F random variable is negative is zero. This must be true since an F random variable is a ratio of two nonnegative numbers. (Y_1/ν_1 and Y_2/ν_2 are both nonnegative.) Once again, it should be emphasised that any F random variable has two numbers of degrees of freedom. Be careful to keep these numbers of degrees of freedom in the correct order, because an F distribution with ν_1 and ν_2 degrees of freedom is not the same as an F distribution with ν_2 and ν_1 degrees of freedom.

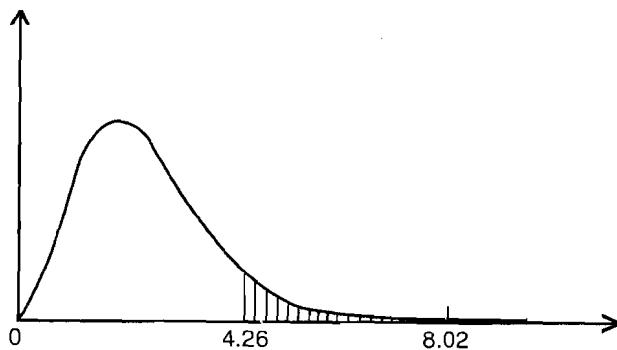


Fig.9

Tables are available which show the values of F that are exceeded with certain probabilities, such as .05 and .01. Appendix Table Q shows, for various numbers of degrees of freedom, the value of F that is exceeded with probability equal to .05. For example, if the numbers of degrees of freedom are 2 and 9, the value of F that is exceeded with probability equal to .05 is 4.26.(See Fig.9.) Similarly, Appendix Table L shows, for various numbers of degrees of freedom, the value of F that is exceeded with probability equal to .01. For example, if the numbers of degrees of freedom are 2 and 9, the value of F exceeded with probability equal to .01 is 8.02.(See Fig.9.)

Let us see some examples.

Example 6: A random variable has F distribution with 40 and 30 degrees of freedom, let us find the probability that it will exceed a) 1.79 b) 2.30.

We first consider the value 1.79. We look at table of F-distribution, Table 4 for 0.05. We proceed downward under column headed ν_2 until entry 30 is reached, then proceed right to the column headed $\nu_1 = 40$ which shows the given value 1.79. Therefore we get that the probability is 0.05.

Since the value cannot be in Table 4 we look at Table 5 which is for 0.01. We proceed similarly and find the value 2.30 in the table. Therefore we get that the probability is 0.01.

* * *

We shall now illustrate the importance of this distribution in the case of sampling distribution.

Suppose that we have two independent random samples of sizes n_1 and n_2 respectively, taken from two normal population having the same variance. If s_1^2 and s_2^2 are sample variances, then the function

$$F = \frac{s_1^2}{s_2^2}$$

has F distribution with v_1 and v_2 degrees of freedom.

F-distribution is used to compare the equality of two sample variances.

Let us consider the following problem.

Problem 6: If two independent random samples of size $n_1 = 7$ and $n_2 = 13$ are taken from a normal population, what is the probability that the variance of the first sample will be at least three times as large as that of the second sample?

Solution: In this we have to compare the variances of the samples. Therefore we apply F-distribution. From Table we find that $F_{.05} = 3.00$ for $v_1 = 7 - 1 = 6$ and $v_2 = 13 - 1 = 12$; thus, the desired probability is 0.05.

————— X —————

Why don't you try these exercises now.

E13) A random variable has the F distribution with 15 and 12 degrees of freedom.

What is the value of this random variable that is exceeded with a probability of 0.5 with a probability of .01?

E14) If a random variable F has 15 and 23 degrees of freedom, what is the probability that it will exceed 2.93.

E15) If two independent random samples of size $n_1 = 9$ and $n_2 = 16$ are taken from a normal population, what is the probability that the variance of the first sample will be at least four times as large as the variance of the second sample?

E16) If independent random samples of size $n_1 = n_2 = 8$ come from normal populations having the same variance, what is the probability that either sample variance will be at least seven times as large as the other?

In this unit you have only been introduced to these three specific distributions t, χ^2 and F. You will realise the utility of these distribution as you study the following units.

This brings us to the end of this unit. We have discussed the sampling distribution of a statistic at length. Let's now briefly recall the various concepts which we have covered here.

4.7 SUMMARY

In this unit we have explained the following concepts

- 1) i) Parameter - Values that describe the characteristic of a population.

- ii) Statistic - Measures describing the characteristic of a sample.
 - iii) Random sample - A sample from a population in which all the items in the population have an equal chance of being in the sample.
 - iv) Sampling distribution of a statistic - For a given population, a probability distribution of all the possible values of a statistic may be taken as for a given sample size.
 - v) Standard error - The standard deviation of the sampling distribution of a statistic.
- 2) We have discussed the central limit theorem.
- 3) We have introduced the following three specific distributions
- i) t-distribution
 - ii) χ^2 -distribution
 - iii) F-distribution
- and explained the use of these distributions as sampling distribution.

4.8 SOLUTIONS/ANSWERS

- E1) Female students enrolled in IGNOU BDP. Finite population
- E2) (ii) and (iii)
- E3) Suppose there are 500 employees working in a company and we are interested to find the average salary paid to them in a month, the average amount calculated from the entire employees is called parameter. On the other hand it is possible to estimate the average based on selected few employees. The average income calculated from sample is called statistic.
- E4) The sampling distribution is given in the following table:

Table 9

Sample No.	Sample elements	Sample Means
1	100, 200, 300	200
2	100, 200, 400	233 (approximately)
3	100, 300, 400	267 (approximately)
4	200, 300, 400	300

The grand mean is 250 and is equal to the population mean.

- E5) i) There are

$$\begin{aligned} {}^6C_2 &= \frac{6 \times 5}{1 \times 2} \\ &= \frac{30}{2} \\ &= 15 \end{aligned}$$

- ii) samples of size 2 are given in the following table.

Table 10

Sample	\bar{X}_i	Sample	\bar{X}_i	Sample	\bar{X}_i
54, 50	52	50, 52	51	52, 50	51
54, 52	53	50, 48	49	52, 52	52
54, 48	51	50, 50	50	48, 50	49
54, 50	52	50, 52	51	48, 52	50
54, 52	53	52, 48	50	50, 52	51

iii) The population mean, the grand mean,

$$\mu = 51.00$$

$$\bar{X} = \frac{\sum \bar{x}_i}{k} = 51.000$$

The grand mean is equal to the population mean.

E6) $N = 200$

$$n = 50$$

$$\sigma = 22$$

$$SE(\bar{X}) = \sigma / \sqrt{n} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$22 / \sqrt{50} \sqrt{\frac{200-50}{199}}$$

$$= 2.7012$$

$$E7) \text{ a) } SE(p) = \sqrt{\frac{250-50}{250-1} \times \frac{\frac{4}{50} \times \frac{46}{50}}{50}} \\ = 0.03438.$$

$$\text{b) } SE(p) = \sqrt{\frac{\frac{7}{60} \times \frac{53}{60}}{60}} \\ = 0.04144.$$

E8) $t_{0.01} = 3.143$

$$t_{0.025} = 2.069$$

$$t_{0.1} = 1.372$$

$$t_{0.005} = 2.756$$

E9) We find the value of the random variable

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

where $\bar{x} = 0.5060$, $\mu = 0.5$ and $s = 0.0040$ and $n = 10$. Substituting, we get the value of t as $t = 4.7434$. From the table we get that the t -value for the parameter $\nu = 9$ and $\alpha = .05$ is 2.26 which is less than the calculate value $t = 4.7434$. Since the probability is less i.e. 0.05, we conclude that the process is out of control.

E10) From the table we find that corresponding to $\alpha = 0.5$ and the parameter

a) $\nu = 16$ the value is 1.75

b) $\nu = 27$ the value is 1.70

c) $\nu = 200$ is 1.645 (note that this value corresponds to the entry to the last row marked ∞).

E11) From the table in the appendix, we find that the value of χ^2_α for $\alpha = 0.5$ is

a) 24.996 for $\nu = 15$ and

b) 32.671 for $\nu = 21$.

E12) We make use of the formula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Note that we are given that $\sigma^2 = 42.5$ and $n = 10$. Let s_1 and s_2 denote the two sample standard deviations, then $s_1^2 = 3.14$ and $s_2^2 = 8.94$ corresponding χ^2 values are

$$\begin{aligned}\chi_1^2 &= \frac{(n-1) \times 3.14}{42.5} \\ &= \frac{9 \times 3.14}{42.5} = 0.664 \\ \text{and } \chi_2^2 &= \frac{9 \times 8.94}{42.5} = 1.893\end{aligned}$$

To find the required approximate probability, it is enough to find the area α_0 given below.

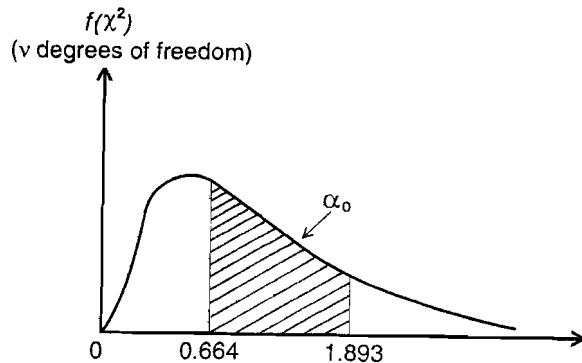


Fig. 10

From the table we find those values of χ^2 for 9 degrees freedom which are close to 0.664 and 1.893.

E13) Table 4 shows that the answer to the first question is 2.62 and Table 5 shows that the answer to the second question is 3.84.

E14) We first look at Table 4 and find that the given value is not there. Then it must be in Table 5 (check it!). Therefore the probability is 0.01.

E15) **Solution** Here $n_1 = 9$ and $n_2 = 16 \therefore \nu_1 = 8$ and $\nu_2 = 15$. We have

$$F = \frac{s_1^2}{s_2^2} = 4$$

To find the probability we look for value $F = 4$ in both the tables and find that it is in Table 5. Therefore the probability is 0.01.

E16) $n_1 = n_2 = 8 \therefore \nu_1 = \nu_2 = 7$ and given that

$$F = \frac{s_1^2}{s_2^2} = 7$$

Proceeding similarly we find the value 6.99 in Table 5 which is almost equal to 7. Therefore the probability is 0.01.

UNIT 5 ESTIMATION

Structure	Page No.
5.1 Introduction Objectives	29
5.2 Point Estimation	30
5.3 Criteria For a Good Estimator	31
5.4 Interval Estimation Confidence Interval for Mean with Known Variance Confidence Interval for Mean with Unknown Variance Confidence Interval for Proportion	34
5.5 Summary	43
5.6 Solutions/Answers	44

5.1 INTRODUCTION

In Units 2 and 3 you have seen that populations can be described by distributions which are fully determined with the help of their parameters. For example, in the case of binomial distribution, you need to know n and p ; in a Poisson distribution, you need to know λ ; and a normal distribution is determined by μ and σ . These quantities are called parameters. The problem with these parameters is that in real-life situations they are usually unknown. We have seen in Unit 4, that in such situations, we take a random sample from the population and compute a function of the sample values, called statistic. More precisely we try to estimate the population parameters by functions of sample values. In this unit we shall discuss certain methods by which we can estimate the population parameters. These processes are called estimation. As we have already stated, in estimation we expect that the sample value is ‘reasonably close’ to the population value. How do you judge this? Here we discuss some criteria which tell us how best the parameters can be estimated by sample values.

In this unit we discuss two methods of estimation - point estimation and interval estimation. In Sec. 5.2 we discuss point estimation. Point estimation concerns choosing a statistic, that is a single number calculated from the sample data. In contrast to this, we sometimes obtain an interval in which we can expect the parameter to lie with some degree of confidence. The method of constructing such intervals is called ‘interval estimation’. In Sec. 5.4, we illustrate construction of such an interval. There we first illustrate how such an interval is constructed for the population mean. We do this in different cases. First we consider the case when the population standard deviation is known and the sample size is large ($n > 30$). Then we take up the case where the standard deviation is unknown, both when the sample size is small and when it is large. After that we shall illustrate how interval estimates are constructed for the population proportion.

Objectives

After reading this unit, you should be able to

- choose an estimator corresponding to a particular situation under study,
- check whether an estimator is,

unbiased
or
efficient.

- construct confidence intervals for the population mean and proportion, using appropriate sampling distribution,
- distinguish between point estimation and interval estimation

5.2 POINT ESTIMATION

Imagine that you need to find the mean life-time of the bulbs produced by a company. Assume that the life of a bulb is distributed as normal with mean θ . Now to find the life-time of a bulb, you have to keep it on till it burns off, and note the time. So, it is a destructive process. If you do this for every bulb produced by the company, it will soon

- have to close down! The way out in this situation is to take a sample of the bulbs, and try to estimate the average life-time of the population on the basis of the life-time observations obtained from the sample. Of course, we cannot hope to get the exact value of the mean life-time. What we get from the sample is only an estimate. If x_1, x_2, \dots, x_n are the life-times of the bulbs which were chosen in a sample of size n , then we could take the sample mean, $\frac{x_1 + x_2 + \dots + x_n}{n}$ as an estimate of the population mean. Of course, this estimate will vary from sample to sample. You already have come across this concept in the previous unit.

But apart from the sample mean, there could be other ways of estimating the population mean from the sample. For example, we could take x_1 as an estimate, or we could take $\frac{x_{\min} + x_{\max}}{2}$ as an estimate where x_{\min} is the minimum value and x_{\max} is the maximum value.

$\frac{x_1 + x_2 + \dots + x_n}{n}$ can be
written as $\frac{\sum_{i=1}^n x_i}{n}$

In any case, the estimate is always based on some or all of the sample values. That is to say that we calculate some sample statistic and take it as an estimate of the population parameter. This sample statistic is called an **estimator**. The value of this estimator for our sample is the **estimate**.

Definition 1: An **estimator** is a function of the sample observations that is used to estimate an unknown parameter. A **point estimate** is a single value of an **estimator**. The process by which we choose an estimator and find the point estimate for estimating an unknown parameter is called **point estimation**.

For example, if a sample mean is used to estimate a population mean, and if the sample mean for a particular sample equals 10, then the estimator used is the sample mean, whereas the point estimate is 10.

To cite another example, suppose we are interested in finding the proportion of individuals in India preferring a given soft drink over another. Here the population parameter is proportion. If the sample proportion is used to estimate the population proportion and if the sample proportion for a particular sample equals 0.6, then the estimator used is the sample proportion and the point estimate is 0.6.

Why don't you try an exercise now.

- E1) Write the estimator and estimate used in the following two situations.

- Suppose an organisation wants to have some information about the mileage for a whole fleet of used taxis, and for that they calculate the mean odometer reading (mileage) from a sample of used taxis and find it to be 98,000 miles.

- ii) Suppose we want to find the proportion of teenagers who have criminal record and for that we take a sample of 50 teenagers and find that 2 % (or .02) have criminal record.
-

We can, in fact, have a number of estimators for a given parameter. Apart from the sample mean, the sample median or the average of the smallest and the largest observations in the sample could also be considered as estimators for the population mean. Since we have a variety of estimators for a parameter θ , we should choose the best of the lot to get a real good estimate. But what do we mean by the best? We'll see that in the next section.

5.3 CRITERIA FOR A GOOD ESTIMATOR

In this section we shall discuss some desirable properties of an estimator. You have already learnt in Unit 4 that an estimator takes different values for different samples. But the estimators such as sample mean, proportion have some nice properties. For example, the sample mean has the property that the means of repeated random sample values taken from a given population will centre on the population mean. You recall that in Unit 4 Sec.4.2, we stated this result that the mean of the sampling distribution of the means is equal to the population mean. This means that 'on the average' the estimator values (or estimates) will equal the parameter value. This property is considered to be one of the criteria for a good estimator. We have a definition here.

Definition 2: Suppose $\hat{\theta}$ (read theta hat) is an estimator of the population parameter, θ . The estimator $\hat{\theta}$ takes different values for different samples. If the mean of all these different estimates is the unknown parameter, θ , then we say that $\hat{\theta}$ is an unbiased estimator of θ . Otherwise, it is called a biased one. It follows from the definition of expectation of a r.v., that $\hat{\theta}$ is unbiased if and only if $E(\hat{\theta}) = \theta$. [Please see Sec.3.2, Unit 3, where we have discussed the expectation of a r.v.]

Let us now look at the estimator given in the following situation.

Let us consider some problems.

Problem 1: A Psychologist measures the reaction times of a sample of 6 individuals to certain stimulus. The measures are given by 0.53, 0.46, 0.50, 0.49, 0.52, 0.53 seconds. Determine an unbiased estimate of the population mean.

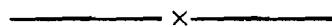
Solution: An unbiased estimate of the population mean is given by the sample mean,

$$\bar{x} = \frac{\sum x_i}{n}$$

Here $n = 6$ and $x_1 = 0.53, x_2 = 0.46, x_3 = 0.50, x_4 = 0.49, x_5 = 0.52, x_6 = 0.53$.

$$\begin{aligned}\bar{x} &= \frac{0.53 + 0.46 + 0.50 + 0.49 + 0.52 + 0.53}{6} \\ &= 0.51 \text{ seconds.}\end{aligned}$$

Then $\bar{x} = 0.5$ seconds. Therefore an unbiased estimate is 0.51 seconds and 0.51 seconds is a point estimate for the mean reaction time of individuals to the stimulus.



Problem 2: In a sample of 400 textile workers, 184 expressed dissatisfaction regarding a prospective plan to modify working conditions. The management felt that this is a strong negative reaction. So they want to know the proportion of total workers who have this feeling of dissatisfaction. Obtain an unbiased estimate of the population proportion.

Solution A point estimate of the population proportion is given by the sample proportion p , given as

$$p = \frac{s}{n}$$

where s denotes the number of observations in the sample which meet the particular characteristic, under study, and n is the sample size.

Here s = 184 and n = 400.

$$\therefore p = \frac{184}{400} = \frac{46}{100}$$

Therefore an unbiased estimate of the population proportion is $\frac{46}{100}$.

Here are some exercises for you.

- E2) A law firm selects a random sample of 60 electronics stores in a particular area, and asks each of them to repair a compact disc player. In each case the law firm determines whether the store makes unnecessary repairs in order to inflate its bill. The law firm finds that 8 of the stores are guilty of this practice. Obtain a point estimate of the proportion of all such stores in the area that inflate bills in this way.
- E3) A washing machine company chooses a random sample of 25 motors from those it receives from one of its suppliers. It determines the length of life of each of the motors. The results (expressed in thousands of hours) are as follows:

4.1	4.6	4.6	4.6	5.1
4.3	4.7	4.6	4.8	4.8
4.5	4.2	5.0	4.4	4.7
4.7	4.1	3.8	4.2	4.6
3.9	4.0	4.4	4.0	4.5

The firm's management is interested in estimating the mean length of life of the motors received from the supplier. Provide a point estimate of this population parameter.

We have seen that the sample mean and sample proportion are unbiased estimates for population mean and population proportion respectively. Does this indicate that the statistic or estimator corresponding to the population parameter is always unbiased? To find an answer to this, let us consider the following example.

Suppose we consider the parameter, 'standard deviation'. Then the sample statistic S given by the formula

$$S = \sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n}} \quad (1)$$

where (x_1, x_2, \dots, x_n) denote the sample observations, can be taken to be an estimator of the population standard deviation. It has been proved that the statistics has an expected value equal to $\sqrt{\left(\frac{n-1}{n}\right)\sigma^2}$ and not σ , this means that **S is not an unbiased estimator of σ** . Hence **an unbiased estimator of σ is obtained by the expression in (2)**

$$\sqrt{\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}}, \quad (2)$$

instead of the expression in (1). For example, an unbiased estimate of the population standard deviation for the situation given in Problem 1 is

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$\begin{aligned}
 \frac{\sum(x_i - \bar{x})^2}{n-1} &= \frac{1}{n-1} [(0.53 - 0.51)^2 + (0.46 - 0.51)^2 + (0.50 - 0.51)^2 + \\
 &\quad (0.49 - 0.51)^2 + (0.52 - 0.51)^2 + (0.53 - 0.51)^2] \\
 &= 0.0006 \\
 \therefore S &= \sqrt{0.0006} \text{ seconds.}
 \end{aligned}$$

As we have seen in E1, in certain situations one can find more than one unbiased estimator for an unknown parameter θ . If we have to choose between two unbiased estimators for a fixed sample size, then we find the standard deviation (or variance) of the sampling distribution of these two estimators and choose that one with smaller standard deviation (or variance). An unbiased estimator T_1 of a parameter θ is said to be more efficient than another unbiased estimator T_2 of θ if $\text{Var}(T_1) \leq \text{Var}(T_2)$, and in such a case, the sampling distribution of T_1 has a smaller dispersion (spread) about θ than that of T_2 (See Fig.1).

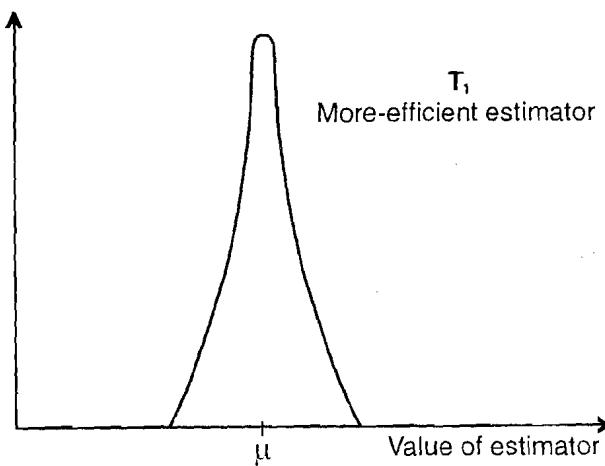
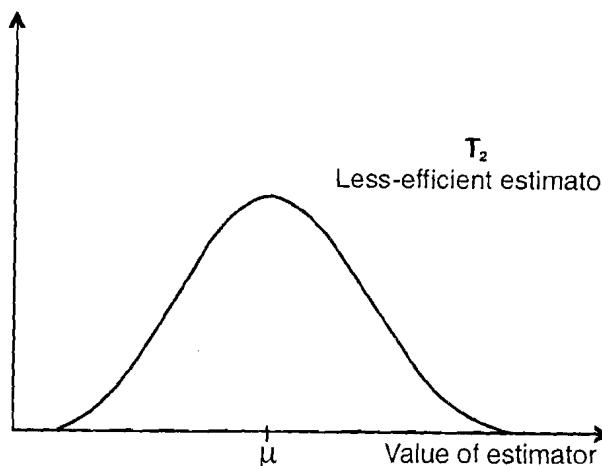


Fig.1:

As an example, let us take a random sample of size n from a normal population with mean μ and standard deviation σ and consider the sample mean and sample median as two estimators of μ . If we compare the sampling distributions of the mean and median for random samples of size n , we get that these two sampling distributions have the same mean but their variances differ. We have seen in Unit 4 that, the variance of the sampling distribution of the mean is σ^2/n , and it can be shown that for random samples of the same size from a normal population, the variance of the sampling distribution of the median is approximately $1.5708 \frac{\sigma^2}{n}$.

Hence we get that both the mean and the median are unbiased estimators, but for a given sample size, the standard error for mean is less than that of median.

From what we have already observed now, we get that for random samples from normal populations the mean is more efficient than the median as an estimator of μ . This fact will be more clear to you when you try E4. In fact it can be shown that in most practical situations where we estimate a population mean μ , the variance of the sampling distribution of no other statistic is less than that of the sampling distribution of the mean. In other words, **in most practical situations the sample mean is the 'most acceptable' statistic for estimating a population mean μ .**

There exist several other criteria for assessing the "goodness" of estimators, but we shall not discuss them in this course.

Why don't you try this exercise now.

-
- E4) To verify the claim that the mean is generally more efficient than the median a student conducted an experiment consisting of 12 tosses of three dice. The following are his results: 2,4, and 6; 5,3, and 5; 4, 5 and 3; 5,2 and 3; 6,1 and 5; 2,3 and 1; 3,1, and 4; 5,5 and 2; 3,3 and 4; 1,6 and 2; 3,3 and 3; and 4,5 and 3.
- Calculate the 12 medians and the 12 means.
 - Group the medians and the means obtained in part (a) into separate distributions having the classes 1.5-2.5, 2.5-3.5, 3.5-4.5 and 4.5-5.5.
 - Draw histograms of the two distributions obtained in part (b) and explain how they illustrate the claim that the mean is generally more efficient than the median.
-

We can summarise our discussion up to now as follows:

- Population parameters are usually unknown and need to be estimated from a sample
 - There could be a variety of estimators for the same parameter.
 - "Unbiasedness" and "efficiency" are some of the desirable properties of a good estimator.
-

5.4 INTERVAL ESTIMATION

In the last section we have seen what a point estimate is. Sometimes it is difficult to evaluate the precision of a point estimator (as measured by its variance, say).

Alternatively, we can think of giving an interval, computed on the basis of the sample values, which will contain the true parameter with a certain degree of confidence. This interval is called an interval estimator of the parameter. These intervals are also called confidence intervals. We shall first discuss confidence intervals for the population mean μ . We first consider the case when the variance σ is known.

5.4.1 Confidence Interval for the Mean with Known Variance

Suppose you have been suspecting that the 1 litre pack of milk that is delivered to your house every morning is not exactly 1 litre, but less. You feel that the filling machine which is supposed to fill each polypack with 1 litre of milk is not working properly. Of course, you are ready to admit that even though the machine is set for 1 litre, it has a certain variability and so there could be some packs which are less than 1 litre full while others which are more.

To end your doubts, you need to find the average volume of milk filled by the machine. Obviously, it would be impossible to do this except by taking a sample. Suppose you

measure the milk pack you get over a period of sixty days. That is, your sample size is 60. Suppose you find that the mean of your observations, which is the sample mean, is 950 ml. This is an estimate of the population mean. But you cannot immediately conclude that the machine is set for 950 ml. You must account for the variability of the sample means. For this you must also know the standard deviation, σ , or calculate it from the sample. Suppose we assume that $\sigma = 50$.

Now we shall construct an interval for the parameter μ the average amount of milk that the machine gives. For that we make use of the central limit theorem discussed in Unit 4. According to this Theorem, for sufficiently large sample size n the sample mean \bar{x} is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Then we make use of the normal distribution table given in Appendix 2 at the end of this block and note that

$$P[-1.96 < Z < 1.96] = 0.95 \quad (3)$$

$$\text{where } Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ i.e. } Z \left(\sigma/\sqrt{n} \right) = \bar{x} - \mu$$

Now we rewrite Eqn.(3) using simple algebra as

$$P \left[-1.96 \frac{\sigma}{\sqrt{n}} < Z \frac{\sigma}{\sqrt{n}} < 1.96 \frac{\sigma}{\sqrt{n}} \right] = 0.95$$

$$\text{i.e. } P \left[-1.96 \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < 1.96 \frac{\sigma}{\sqrt{n}} \right].$$

Now we subtract $-\bar{x}$ from all the three terms inside the bracket. Then we get

$$P \left[-\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right] = 0.95$$

Now we multiply all the terms inside the bracket by -1 and (therefore the inequalities get reversed) and we get

$$P \left[\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right] = 0.95 \quad (4)$$

Thus corresponding to each sample mean \bar{x} , we got an interval given by

$$\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \quad (5)$$

which satisfies Equation (4). Let us now see what does Equation (4) implies. Let us, for example consider the sample value $\bar{x} = 950$ ml. obtained for the problem regarding average volume of milk filled by the machine. Then the Equation (4) corresponding to $\bar{x} = 950$ ml is

$$P[937.35 < \mu < 962.65] = 0.95$$

We interpret it in the way that we are 95 % confident that the interval (937.35, 962.65) contains the true value μ . This does not mean that "There is 95 % probability that μ lies in the interval (937.35, 962.65). This is a very common mis-interpretation of Equation (4) and it is incorrect. This is because the population mean μ is a fixed quantity and therefore μ either lies in the interval (937.35, 962.65) or it does not. Therefore the probability that μ lies in the interval is either 0 or 1. The 95 percent probability is assigned to our level of confidence that the interval contains μ . It is not assigned to the probability that μ lies in the interval.

Another interpretation of Equation (4) is based on the fact that we can construct a confidence interval for each sample mean \bar{x} . We will get different intervals for different values of sample means. So, in this case Equation (4) says that if all possible samples of size n are calculated, and the intervals $\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$ are calculated for each sample, then 95 % of all such intervals are expected to contain the population parameter μ . This does not mean that for a particular sample value \bar{x} , we can expect that the interval $\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$ will contain μ .

If we multiply the terms in the inequality $y \geq 1$ by (-1), then the inequality gets reversed and we get $-y \leq -1$.

The confidence intervals $\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$ is also denoted as $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

That means if you select 100 samples and calculate the intervals about their sample means, then 95 of these will contain the population μ . Note that here we assume that σ is known. In the following figure we have illustrated this graphically, showing five such intervals

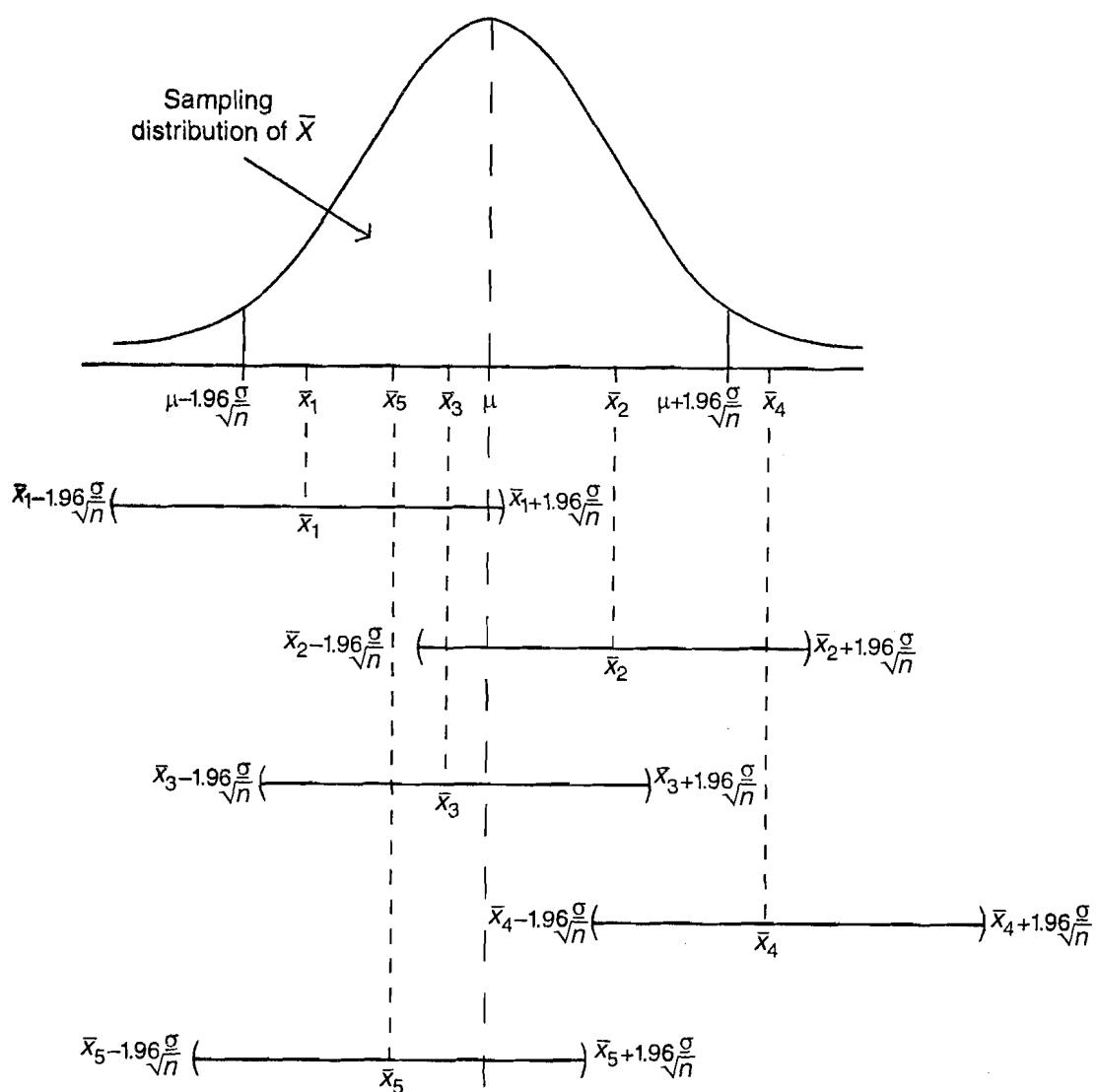


Fig.2: A number of intervals constructed around the population mean.

Only the interval constructed around the sample mean \bar{X}_4 does not contain the population mean.

The interval given by (5) is called a confidence interval.(C.I)

The value 0.95 (or 95%) attached with the confidence interval is called **confidence coefficient**. The left end point of the confidence interval is called **lower confidence limit (LCL)** and the right end point of the confidence interval is called **upper confidence limit (UCL)**. The difference between the UCL and LCL is the width of the confidence interval. The width of the 95% confidence interval in the above example is $2 \times \left(1.96 \frac{\sigma}{\sqrt{n}} \right)$

Although 0.95 is frequently used as a confidence coefficient, we can have other values such as 0.90 or 0.99 as confidence coefficients. Using the normal distribution table, we can obtain the confidence interval for 0.90 (or 90%) as

$$\left(\bar{x} - 1.64 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.64 \frac{\sigma}{\sqrt{n}} \right)$$

and for 0.99 (or 99%) as

$$\left(\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}} \right)$$

In the following problem we illustrate the use of confidence intervals.

Example 1: The Director of a marketing division wants to analyse the market value of business firms of a similar size. [Market value is defined as the number of common shares outstanding, multiplied by the share prize as listed on an organised exchange]. A sample of 600 firms revealed a mean market value of Rs.850 million. The earlier results reveals that the population is normally distributed with population standard deviation Rs.200 million. It is desired to set up a confidence interval for the (unknown) mean market value.

Given that the sample mean is $\bar{x} = 850$ million and the sample size $n = 600$ and $\sigma = 200$ million. Therefore can construct 95 % confident interval, which is given by

$$\left(850 - 1.96 \times \frac{200}{\sqrt{600}}, 850 + 1.96 \times \frac{200}{\sqrt{600}} \right)$$

i.e. $(850 - 15.99, 850 + 15.99)$
i.e. $(834.01, 865.99)$

This shows that the director can be 95% confident that the interval (834.01, 865.99) contain the mean market value.

* * *

It is time to do some exercises.

- E5) For each of the values given below, calculate the 95% confidence interval for the mean.
- $\bar{x} = 0, \sigma = 10, n = 8$
 - $\bar{x} = 550, \sigma = 40, n = 16$.
- E6) If the mean length of hospitalisation of 140 patients was 11.4 days and the standard deviation of patient days is assumed to be 2.5 days, what is the 99% confidence interval for the average length of stay? Assume normality.
- E7) Estimate the number of days between germination and the first pickable cucumbers using the following sample.

Date of germination	First Fruit
May 1	June 17
4	18
8	21
5	16
12	28
18	July 3
11	June 25
9	26

What is the 95% confidence interval assuming $\sigma = 2$ days?

Next we shall consider the case when σ is unknown.

5.4.2 Confidence Interval for Mean with Unknown Variance

In all the computations of the confidence interval for μ so far we have assumed that the population variance is known. Each time, the normal distribution was the appropriate sampling distribution used to determine the confidence intervals. However the normal

distribution is not appropriate when the population variance is unknown and the sample size is less than 30. In such situations we use t-distribution. As indicated in the previous unit, the sample standard deviation 's' is generally used as an estimator of the population standard deviation.

If the sample size is 30 or less and the population is normal (and large relative to the sample), a confidence interval for the population mean can be constructed by using the t-distribution in place of the standard normal distribution.

You are already familiar with t distribution from Unit 4. We now have to use the t distribution table given in Appendix to construct the confidence intervals corresponding to different levels of confidence, say 95% or 99%. Let us suppose that we want to find the confidence intervals at the 90% confidence level i.e. $\alpha = 0.1$ with a sample size of 14 similar to the ones we have given in Equation(5). Note that we don't know σ in this case. Therefore, as indicated in Unit 4, the sample standard deviations is used as an estimator of the population standard deviation. Thus if s is known, then 90% confidence interval is given as

$$\left(\bar{x} - t_{0.05} \frac{s}{\sqrt{n}}, \bar{x} + t_{0.05} \frac{s}{\sqrt{n}} \right)$$

where $t_{0.05}$ is the t-value corresponding to the value $\frac{\alpha}{2} = \frac{0.1}{2} = 0.05$ and for the parameter $\nu = n - 1$, where n is the sample size. Now to find the t-value we make use of the table 1 in the Appendix. For example, suppose that $n = 14$, then $\nu = 13$, then, from table 1 we get that the t-value is $t_{\alpha/2} = 1.771$ (See Fig. 3).

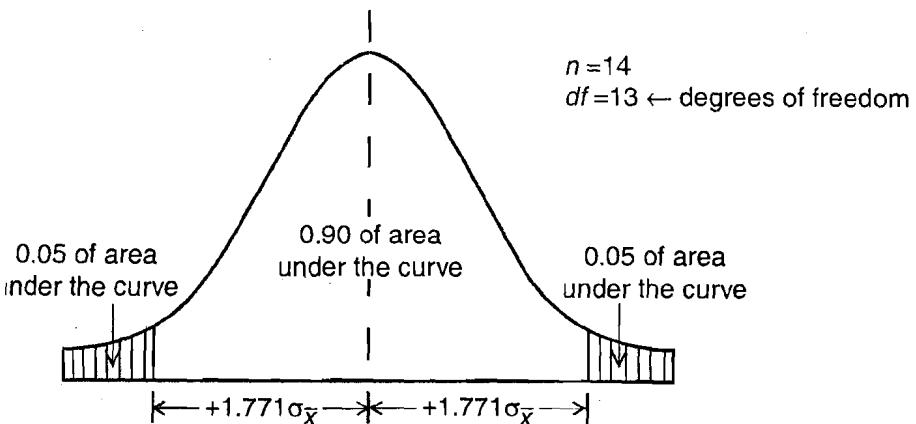


Fig.3: Confidence interval using the t-distribution

Like a z-value, the t-value 1.771 shows that if we mark off plus and minus 1.771 $\frac{s}{\sqrt{n}}$ on either side of the mean \bar{x} , the area under the curve between these two limits will be 90%, and the area outside these limits will be 10%.

Therefore the **90% confidence interval, for degrees of freedom 13 is.**

$$\left(\bar{x} - 1.771 \frac{s}{\sqrt{n}}, \bar{x} + 1.771 \frac{s}{\sqrt{n}} \right) \quad (6)$$

Similarly the **95% confidence interval for 20 degrees of freedom is**

$$\left(\bar{x} - 2.086 \frac{s}{\sqrt{n}}, \bar{x} + 2.086 \frac{s}{\sqrt{n}} \right) \quad (7)$$

In a similar way we can find), confidence intervals for different degrees of freedom. (see E8)

Let us consider some examples.

Example 2: A sample of 10 measurements of the diameter of a sphere has a mean $\bar{x} = 43.5\text{mm}$ and $s^2 = 4\text{mm}^2$. Let us find the i) 95% and ii) 99% confidence intervals for the actual diameter.

- i) Here $n=10$ and $\alpha = 0.5$. Therefore, we use t distribution with 9 d.f. From the table, we get $t_{\alpha/2} = t_{0.025} = 2.26$. So, the 95% confidence interval for μ is

$$\begin{aligned} & \left(\bar{x} - t_{0.025} \left(\frac{s}{\sqrt{n}} \right), \bar{x} + t_{0.025} \left(\frac{s}{\sqrt{n}} \right) \right) = \\ & \left(43.5 - 2.26 \left(\frac{2}{\sqrt{10}} \right), 43.5 + 2.26 \left(\frac{2}{\sqrt{10}} \right) \right) = (42.07, 44.93). \end{aligned}$$

So, we can be 95% confident that the true mean lies between 42.07 and 44.93

- ii) Working similarly, the 99% confidence interval is
 $(43.5 - 3.25 \left(\frac{2}{\sqrt{10}} \right), 43.5 + 3.25 \left(\frac{2}{\sqrt{10}} \right)) = (41.44, 45.55).$

* * *

The idea will be more clear to you if when you do the following exercises.

Problem 3: A manufacturer of light bulbs wants to estimate the mean length of life of a new type of bulb which is designed to be extremely durable. The firm's engineer tests nine of these bulbs and find that the length of life (in hours) of each is as follows:

5,000 5,100 5,400
 5,200 5,400 5,000
 5,300 5,200 5,200

Previous experience indicates that the lengths of life of individual bulbs of a particular type are normally distributed. Construct a 90 percent confidence interval for the mean length of life of all bulbs of this new type.

Solution: If x_i is the length of life of the i th light bulb in the sample, we find that

$$\begin{aligned} \sum_{i=1}^9 x_i &= 46,800 \\ \bar{x} &= 5,200 \\ \sum_{i=1}^9 (x_i - \bar{x})^2 &= 1,80,000 \\ \sum_{i=1}^9 (x_i - \bar{x})^2 / (n - 1) &= 22,500 \\ \therefore s &= \sqrt{22,500} = 150. \end{aligned}$$

Since $n=9$, we make use of t-distribution. Because a 90 percent confidence interval is wanted, $t_{\alpha/2} = t_{0.05}$; and the number of degrees of freedom is $(n - 1) = 8$. Therefore, the t-distribution table given in the Appendix of Unit 4 shows that if there are 8 degree of freedom, $t_{0.05} = 1.86$. Thus, the desired confidence interval is

$$5200 \pm 1.86 \left(\frac{1.50}{\sqrt{9}} \right) \quad \left(5200 - 1.86 \left(\frac{1.50}{\sqrt{9}} \right), 5200 + 1.86 \left(\frac{1.50}{\sqrt{9}} \right) \right)$$

By simplifying, we get that the confidence interval is (5107, 5293).

————— X —————

Why don't you try these exercises now?

- E8) Given the following sample sizes and confidence levels, find the appropriate $t_{\alpha/2}$ values for constructing confidence intervals.
- i) $n = 10$; 99%
 - ii) $n = 28$; 95%
 - iii) $n = 13$; 90%

iv) $n = 25$; 99%

- E9) A sample of 12 measurements of breaking strengths of cotton threads gave a mean of 0.738 N and a standard deviation of 0.124N. Find a 95% and 99% confidence intervals for the actual breaking strength.
- E10) Five measurements of the reaction time of an individual to certain stimuli were recorded as: 0.28, 0.30, 0.27, 0.33 and 0.31 second. Find the 95% confidence interval for the actual reaction time.
- E11) If you are given a sample of 20 candles from a large shipment of candles, and are asked to give an interval estimate of their average burning life, how would you proceed? What information would you need?
-

The above examples and exercises illustrate how we can use t-distribution to find the confidence intervals. As we mentioned earlier, **t-distribution can be used only if the population variance is unknown and the sample size is small**. Next we shall see how to construct the confidence intervals for large samples when the population variance is unknown.

Mathematicians have shown that if the sample size is large, we can simply substitute the sample standard deviation for the population standard deviation in the results obtained in the previous part of this section i.e. in Subsection 5.4.1. Thus, if we want to construct a 95% confidence interval — that is, a confidence interval with a confidence coefficient of 95 percent — we can substitute s for σ in Equation (5), the result being

$$p_r \left[\bar{x} - 1.96 \frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{s}{\sqrt{n}} \right] = .95 \quad (8)$$

Consequently, the confidence interval is

$$\left(\bar{x} - 1.96 \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \frac{s}{\sqrt{n}} \right) \quad (9)$$

Equation (8) is applicable only if the population is large relative to the sample.

The following example should make the above discussion more clear.

Problem 4: A random sample of 100 ball bearings made by a machine in 1 week was taken. The mean diameter was found to be 8.24 mm with a standard deviation of 0.42 mm. Find the 95% and 99% confidence intervals for the mean diameter of ball bearings produced by that machine.

Solution: Since the sample is large, from Equation(8), we get that the 95% confidence interval for μ is

$$\left(8.24 - 1.96 \left(\frac{0.42}{\sqrt{100}} \right) \right), \left(8.24 + 1.96 \left(\frac{0.42}{\sqrt{100}} \right) \right) = (8.16, 8.32)$$

Similarly, the 99% confidence interval is

$$\left(8.24 - 2.58 \left(\frac{0.42}{\sqrt{100}} \right) \right), \left(8.24 + 2.58 \left(\frac{0.42}{\sqrt{100}} \right) \right) = (8.13, 8.35)$$

See if you can solve these exercises:

-
- E12) A random sample of marks obtained by 50 students in Mathematics showed a mean of 75 and a standard deviation of 10.
- What are the 95% confidence limits for the mean marks in Mathematics?
 - With what degree of confidence can we say that the mean marks are between 74 and 76?

E13) A washing machine company's statistician says that 90% confidence interval for the mean length of motors received from Supplier II is 4,500 to 4,800 hours, based on a sample of 36 motors. The statistician also says that the standard deviation of the lengths of life of motors received from Supplier II is 500 hours. Is there any contradiction between the statements? If so, what is the contradiction?

Next we shall illustrate how confidence intervals are calculated for population proportions. We have talked in length about the estimation of population parameter μ . Another important population that we need to estimate is the population proportion, p . Let's see how to go about it.

5.4.3 Confidence Interval for Population Proportion

Let us start with a situation. In a random sample of 25 men from a city, 8 were found to be smokers. Can we estimate the proportion of smokers in the city?

Suppose the proportion of smokers in a city is π . Then $p = \frac{8}{25}$ is a point estimate of π , obtained from this sample. Now we shall construct a confidence interval for estimate π . To do this we proceed similar to what we did for the population mean. We recall the result from Unit 4, that the sampling distribution of sample proportion p has mean π and

standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$. We also know from Unit 4, that if the sample size is sufficiently large and if π is not very close to 0 or 1, the sampling distribution is approximately a normally distribution. Then using the standard normal distribution table, we can find confidence intervals. If we want to construct 95% confidence intervals then that will be given by

$$\left(p - 1.96 \sqrt{\frac{\pi(1-\pi)}{n}}, \quad p + 1.96 \sqrt{\frac{\pi(1-\pi)}{n}} \right) \quad (10)$$

so that

$$P \left[p - 1.96 \sqrt{\frac{\pi(1-\pi)}{n}} < \pi < p + 1.96 \sqrt{\frac{\pi(1-\pi)}{n}} \right] = 0.95$$

The interval given by in (9) is called 95% confidence interval for π . Similarly, we can have 90% or 99% confidence intervals. The above intervals given in Equation (9) cannot be used as they involve the unknown, π .

However, if n is large, then π can be replaced by p without compromising accuracy. So that for large samples, the 95% confidence interval for π will be

$$\left(p - 1.96 \sqrt{\frac{p(1-p)}{n}}, \quad p + 1.96 \sqrt{\frac{p(1-p)}{n}} \right)$$

If we want to get a 99% confidence interval, we will have to replace 1.96 by 2.58, since

$$P \left[-2.58 \leq \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} \leq 2.58 \right] = 0.99$$

We shall illustrate this with the problem given below.

Problem 5: In a random sample of 75 parts produced by a machine, 12 have a surface finish which is rougher than the specification will allow. Find a i) 95% ii) 99% confidence interval for the proportion of rough parts produced by the machine.

Solution: Here $p = 12/75 = 0.16$. Then

a) 95% confidence interval is

$$\left(0.16 - 1.96 \sqrt{\frac{0.16 \times (0.84)}{75}}, \quad 0.16 + 1.96 \sqrt{\frac{0.16 \times (0.84)}{75}} \right) = (0.08, 0.24).$$

b) The 99% confidence interval for P is

$$\left(0.16 - 2.58 \sqrt{\frac{.16 \times (.84)}{75}}, 0.16 + 2.58 \sqrt{\frac{.16 \times (.84)}{75}} \right) = (0.05, 0.27)$$

Here are some exercises for you.

E14) A random sample of 800 calculators contains 24 defective items. Compute a 99% confidence interval for the proportion of defective calculators.

E15) Of 1000 randomly selected lung cancer cases, 699 resulted in death. Construct a 95% confidence interval for the death rate from lung cancer.

E16) A student in a university wanted to decide whether or not a contest the election for the presidency of the students' union. Out of 50 students, 11 showed their willingness to vote for her. Find a 99% confidence interval for the true proportion of students voting for her.

We now summarise our discussion about interval estimation in the following table:

Table 1

Parameter		Point Estimator	Confidence Interval
μ	$\bar{x} \longrightarrow$	σ known	$(\bar{x} - z \frac{\sigma}{\sqrt{n}}, \bar{x} + z \frac{\sigma}{\sqrt{n}})$
		σ unknown, large n	$(\bar{x} - z \frac{s}{\sqrt{n}}, \bar{x} + z \frac{s}{\sqrt{n}})$
		σ unknown, small n	$(\bar{x} - t \frac{s}{\sqrt{n}}, \bar{x} + t \frac{s}{\sqrt{n}})$
π	p	p is not too close to 0 or 1, large n	$(p - z \sqrt{\frac{p(1-p)}{n}}, p + z \sqrt{\frac{p(1-p)}{n}})$

Now we shall present a case study which shows how the techniques of the estimation discussed in this unit helps in tackling real-life problems.

Abrasion resistance of rubber is the extent to which rubber can withstand pressure against rubbing off or frictional action. For example, rubber of high abrasion resistance will have high road life.

Statistical Estimation in the Chemical Industry: A Case Study: A chemical firm called Imperial Chemical Industry (ICI) carried out the following experiment to estimate the effect of a chlorinating agent on the abrasion resistance of a certain type of rubber. Ten pieces of this type of rubber were cut in half, and one half-piece was treated the chlorinating agent, while the other half-piece was untreated. Then the abrasion resistance of each half was evaluated on a machine, and the difference between the abrasion resistance of the treated half-piece and the untreated half-piece was computed. Table below shows the 10 differences (1 corresponding to each of the pieces of rubber in the sample). Based on this experiment, ICI was interested in estimating the mean difference between the abrasion resistance of a treated and untreated half-piece of this type of rubber. In other words, if this experiment were performed again and again, an infinite population of such differences would result. ICI was interested in estimating the mean of this population, since the mean is a good measure to find the effect of the chlorinating agent on this type of rubber's abrasion resistance.

If you were a statistical consultant for ICI, how would you analyse these data? You would recognise that a good point estimate of the mean of this population is the sample mean, which is 1.27 as shown in Table below. Thus, your first step would be to advise ICI that if they want a single number as an estimate, 1.27 is a good number to use. Next you would point out that such a point estimate contains no indication of how much error it may contain, whereas a confidence interval does contain such information. Since the population standard deviation is unknown and the sample is small, expression () should be used in this case to calculate a confidence interval. Assuming that the firm

wants a confidence coefficient of 95 percent, the confidence interval is (0.464 – 2.076), because $t_{0.025} = 2.262$, $s = 1.1265$, and $n=10$. The chances are 95 out of 100 that such a confidence interval would include the population mean. (Note that this analysis assumes that the population is approximately normally distributed)

Place	Difference
1	2.6
2	3.1
3	-0.2
4	1.7
5	0.6
6	1.2
7	2.2
8	1.1
9	-0.2
10	0.6

$$\sum_{i=1}^{10} x_i = 12.7$$

$$\bar{x} = 1.27$$

$$s = 1.1265$$

The above analysis is, in fact, exactly how ICI's statisticians proceeded. Despite the fact that the sample consisted of only 10 observations, the evidence was very strong that the chlorinating agent had a positive effect on abrasion resistance. After all, the 95 percent confidence interval was that the mean difference between abrasion resistance of rubber with and without treatment was an increase of between 0.464 and 2.076. (For that matter, the statisticians found that the 98 percent confidence interval was that the mean difference was an increase of between 0.265 and 2.275). The best estimate was that the chlorinating agent resulted in an increase of about 1.27 in abrasion resistance.

With the detailed example you have seen how several aspects covered in this unit has merged. In fact as you reflect on this case study you should check from the summary below how many points are actually covered in this case study.

With that we come to the end of this unit.

5.5 SUMMARY

In this unit we have seen that

- 1) When the population is large, its parameters, like mean, variance, proportion, need to be estimated from a sample.
- 2) There can be many different estimates of a parameter.
- 3) An estimator is unbiased if the mean of the estimates is the population parameter
- 4) Between two unbiased estimators we prefer the one with the smaller variance
- 5) Interval estimates are better than point estimates since we can easily specify the precision of our estimate.
- 6) The computation of confidence intervals is done by using the sampling distributions of the estimators.

5.6 SOLUTIONS/ANSWERS

- E1) For (i) the estimator is the mean mileage of the sample of used taxis. The value 98,000 miles is an estimate.

For (ii) the estimator is the proportion and the value .02 is an estimate.

- E2) An unbiased estimate of the population proportion is obtained by

$$p = \frac{8}{60} = \frac{2}{15}$$

- E3) A point estimator of the population mean is obtained by calculating the sample mean.

The sample mean of 25 motors is 4.448 thousands of hours.

- E4) a) The medians are 4,5,4,3,25,2,3,5,3,2,3 and 4; the means are 4,4.3, 4,3.3,2, 2.7,4,3.3,3 and 4.

b) The frequencies are 2,4,3 and 3 for the medians and 1,5,6 and 0 for the means. Then obtain the frequency distribution.

c) The histograms of two distributions shows that the variance for the median is more than for the mean which illustrate the claim that the mean is generally more efficient than the median.

- E5) 95% confidence interval for mean is $\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$

i) Here $\bar{x} = 0$ and $\sigma = 10$ and $n = 8$. Therefore the interval is

$$\left(0 - 1.96 \frac{10}{\sqrt{8}}, 0 + 1.96 \frac{10}{\sqrt{8}} \right) = (-6.9296, 6.9296)$$

ii) Here $\bar{x} = 550$, $\sigma = 40$ and $n = 16$. Therefore the interval is

$$\left(550 - 1.96 \frac{40}{\sqrt{16}}, 550 + 1.96 \frac{40}{\sqrt{16}} \right) = (530.4, 569.6)$$

- E6) Here $n = 140$, $\bar{x} = 11.4$, $\sigma = 2.5$

99% C.I. is given by

$$\left(\bar{x} - 2.58 \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}} \right) = \left(11.4 - 2.58 \frac{2.5}{\sqrt{140}}, 11.4 + 2.58 \frac{2.5}{\sqrt{140}} \right) = (10.855, 11.945)$$

- E7) Number of days are:

47,45,44,42,47,46,45,48

$\therefore \bar{x} = 5.5$, $n = 8$, $\sigma = 2$

There C.I. is given by $\left(5.5 - 1.96 \frac{2}{\sqrt{8}}, 5.5 + 1.96 \frac{2}{\sqrt{8}} \right) = (4.1141, 6.8859)$

- E8) i) Note that here $t = n - 1 = 10 - 1 = 9$ and $\alpha = 0.005$. Therefore we look under column for 0.005 till we reach the row for 9. Then we get the value 3.250.

ii) Here $\alpha = 0.5$ the $t_{\alpha/2}$ -value is 2.052

iii) Here $\alpha = 0.1$ the $t_{\alpha/2}$ -value is 1.782

iv) Here $\alpha = 0.01$ the $t_{\alpha/2}$ -value is 3.797

- E9) $n = 12 \therefore d.f. = 11$

The t value for 95% C.I. is 2.20 and that for 99% C.I. is 3.11.

$$\therefore 95\% \text{ C.I.} : 0.738 \pm 2.20 \left(\frac{0.124}{\sqrt{12}} \right) = (0.6592, 0.8167)$$

$$99\% \text{ C.I.} : 0.738 \pm 3.11 \left(\frac{0.124}{\sqrt{12}} \right) = (0.6267, 0.8493)$$

E10) From the sample, $\bar{x} = 0.298$ and

$$s = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} = 0.0213 \text{ and d.f.} = 4.$$

\therefore The required value of t is 2.78

$$\therefore 95\% \text{ C.I.} = 0.298 \pm 2.78 \left(\frac{0.0213}{\sqrt{5}} \right)$$

$$= (0.2715, 0.3245)$$

E11) Light up the candles and measure the amount of time (life time) for which each candle burns. This data will have 20 observations. Find the mean (\bar{x}) and the standard deviation (s) of this data. The value of \bar{x} is a point estimate. If we want 95% C.I., we find $t = 2.09$ for 19 d.f., since the sample size is 20. Then C.I. is $\bar{x} \pm \frac{s}{\sqrt{20}}$.

E12) a) (72.2281, 77.7718)

$$\begin{aligned} b) P(74 \leq \mu \leq 76) &= 2P(75 \leq \mu \leq 76) \\ &= 2P\left(0 \leq Z \leq \frac{76.75}{10/\sqrt{5}}\right) \\ &= 2P(0 \leq Z \leq 0.7071) = 0.5224 \end{aligned}$$

\therefore Degree of confidence is 52%

E13) C.I. is $p \pm 2.58 \sqrt{\frac{p(1-p)}{n}}$

$$p = \frac{24}{800} = 0.03, n = 800$$

$$\therefore \text{C.I.} = (0.0144, 0.0456)$$

E14) $p = 0.699$

$$\text{C.I. is } (0.6706, 0.7274)$$

E15) $p = \frac{11}{50} = 0.22$. Therefore the C.I is $(0.0688, 0.3711)$

UNIT 6 TESTS OF SIGNIFICANCE

Structure	Page No.
6.1 Introduction Objectives	46
6.2 Some Basic Concepts	47
6.3 Tests About the Mean Difference in the Means of Two Populations	53
6.4 Test About the Variance	66
6.5 Tests About the Population Proportion	67
6.6 Summary	69
6.7 Solutions/Answers	69

6.1 INTRODUCTION

Statistical data are collected in many diverse fields. The main purpose of data collection is to arrive at certain decisions. For example, a company may want to decide whether or not to accept a consignment of ball bearings. A student may want to know whether or not to believe the claim of a coaching class owner that students in his class get 90% marks on the average in Maths. A doctor may need to test whether or not to prefer the new pain medicine to the old.

In all these situations one can identify a random variable whose distribution is not known to the decision maker, but will be useful for decision making. The company may identify a quality characteristic associated with each ball bearing and the company may want to know if the quality characteristic associated with a great majority of the ball bearings in the consignment satisfy the upper and lower specifications on it. The student wants to know if the expected value of the random variable, namely the score in Mathematics of a student of the coaching centre is 90% or not. The doctor wants to know whether the expected value of the random variable, which is the ‘duration of relief’ for the new pain medicine is more than that for the old medicine or not. In each of the above cases the decision has to be based on the information obtained from a random sample. In the previous unit you have seen how we can get a point and an interval estimates of parameters of a distribution by observing a finite number of realizations of the random variable. In this unit we shall see how a hunch or a claim or a hypothesis about the actual value of a parameter can be tested and a decision taken to reject or accept the hypothesis.

Let us consider again the random variables that were of interest in the above mentioned situations. For the company that has received a consignment of ball bearings, the quality characteristic of interest may be its thickness. The thickness of a ball bearing can be measured on a continuous scale and if X_1, X_2, \dots, X_n denote n realizations (the thickness of n randomly chosen ball bearings, from a very large consignment), it is easy to see that they are a set of independent and identically distributed random variables. It is reasonable to assume that the common distribution is the normal distribution with some unknown mean μ and variance σ^2 . The student can reasonably assume that the scores in Mathematics of a random sample of size n from among the alumni of the coaching centre are independent and identically distributed as normal with mean μ and variance σ^2 . Similarly, the time durations of relief reported by n_1 randomly chosen test

patients using one dose of the new pain medicine can be modelled as n independent and identically distributed random variables distributed as normal(μ_1, σ_1^2) and those of another set of randomly chosen n_2 patients using one dose of the old medicine as i.i.d. normal(μ_2, σ_2^2). We shall deal with the problem of testing hypotheses about the mean and variance of normally distributed random variables in sections 6.3 and 6.4. In Sec. 6.5 we shall illustrate the problem of testing hypothesis about the proportion.

Objectives

After reading this unit, you should be able to

- formulate null and alternative hypothesis for a given problem,
- describe Type 1 and Type 2 errors,
- differentiate between one sided alternatives and 2-sided alternatives,
- test whether μ has the specified value μ_0 against either a one sided alternative or against a two sided alternative assuming that we have n independent observations on a random variable that is distributed as normal(μ, σ^2) where σ is either assumed known or is unknown,
- test whether the two normal populations have the same mean or not,
- test whether the population variance from a normally distributed population, has a specified value against either a one sided or two sided alternative,
- test if the population proportion is significantly different from the hypothetical population proportion,
- test if the difference between two population proportions is zero or not based on the observed sample proportions.

6.2 SOME BASIC CONCEPTS

Let us begin with a problem.

A company manufacturing Aluminium rods wants to find the average mean diameter of the length of rods manufactured in the company. Based on previous experience, the firm's statistician knew that, on the average, the rods should be 2 cm. in diameter. He also knew that the variance of the diameters of the rods produced by the process is $\sigma_0^2 = 0.25$. A test was needed for detecting any change in this mean diameter every day so that whatever factors were responsible for such a change could be corrected.

Note that in this situation the statistician is interested in making inference about the population parameter mean μ . However, she is not interested in estimating the value of μ , rather she is interested in testing a hypothesis about the value of μ . The hypothesis is that the mean diameter of the rods produced on a particular day is 2cm. (That is there has been no change in the mean diameter (μ) of all the rods produced in a day) from the established standard length. Let us denote this hypothesis by H_0 , and call it the null (no change) hypothesis. The statistician wishes to test this null hypothesis against the alternative that it is not true.

In statistics, the **null hypothesis is the hypothesis that is being tested**. A statistician has to be careful while formulating the null and alternative hypotheses since it is presumed that the null hypothesis holds unless there is a strong statistical evidence obtained from the sample of observations against it and in favour of the alternative hypothesis. The situation is somewhat similar to the modern system of justice. A person accused of a crime is presumed to be not guilty, unless the prosecution can produce strong evidence to the contrary. To make the concept clear, let me pose a question. A

strong evidence to the contrary. To make the concept clear, let me pose a question. A manufacturer of paints wants to use a new drying process. The present process requires a drying time of 15 minutes and it is claimed that the new process requires only 12 minutes of drying. The new process is to be tested on n sample pieces. What should be the null and alternative hypothesis if the manufacturer wants to continue to use the old process unless there is strong evidence for the claim regarding the new process? In this case you can see that the null hypothesis is $H_0 : \mu = \mu_0 = 15$ minutes and the alternative hypothesis is $H_1 : \mu = \mu_1 = 12$ minutes, where μ is the mean drying time of the new process. The new process is not accepted unless the null hypothesis is rejected. This is the case of a conservative manufacturer. A more risk taking manufacturer may decide that the new process may be accepted unless there is statistical evidence against the claim. In this case the null hypothesis is formulated as $H_0 : \mu = \mu_1 = 12$ minutes against the alternative hypothesis $H_1 : \mu = \mu_0 = 15$ minutes. Unless the null hypothesis in this formulation is rejected, the manufacturer goes for the new process.

We now continue with the case of the company manufacturing aluminium rods. The statistician is investigating if the diameter of a rod produced today is distributed as normal with a mean that is different from 2.0 cm. Unless there is strong statistical evidence for such a change, the statistician will not like to reject the hypothesis that $\mu = 2$ as that might entail stopping the process. If such a change is absent, the mean diameter is 2 cm. Thus the null hypothesis H_0 in this case is

$$H_0 : \text{The mean diameter is } 2 \text{ cm.}$$

Now to make any decision, the company have to choose between this hypothesis and the alternative hypothesis, which we denote H_1 , and this is stated as

$$H_1 : \text{The mean diameter of the given lot of rods is not } 2 \text{ cm.}$$

As the name suggests this hypothesis is alternate to H_0 .

Thus the statistician has made a claim/hypothesis and she wants to test this claim against a suitable alternative hypothesis on the basis of a sample of observations.

Suppose X_1, X_2, \dots, X_n are i.i.d. (independently and identically distributed) random variables having the common distribution function $F(\theta)$ where θ is a real valued parameter. The set of all possible values of θ is known as the parameter space and is denoted as Θ . We wish to test the hypothesis that $\theta \in \Theta_1$ against the alternative that $\theta \in \Theta_2$, where Θ_1 and Θ_2 are non-intersecting subsets of Θ . In the above example, the statistician will observe from the day's production a random sample of n rods. Let X_1, \dots, X_n denote their diameters. Clearly, X_1, X_2, \dots, X_n are i.i.d random variables. The statistician also assumes that their common distribution is normal with mean μ and variance σ_0^2 , where σ_0^2 is assumed to be 0.25. The parameter space Θ may be taken as $(-\infty, \infty)$ although the mean diameter can never be negative. Also we take $\Theta_1 = \{2.0\}$ and $\Theta_2 = \Theta \setminus \{2.0\}$.

We call the null hypothesis *simple* if Θ_1 is a singleton set. Otherwise we call it a *composite hypothesis*. Similarly, the alternative hypothesis is called simple if the set Θ_2 is a singleton set; else it is called a composite hypothesis. The testing problem of the company manufacturing aluminium rods has a simple null hypothesis against a two sided (i.e., both $\mu > \mu_0$ and $\mu < \mu_0$) composite alternative hypothesis.

In any hypothesis testing problem such as the above you may easily recognise four possibilities, two for the true value of the hypothesis H_0 and two possibilities for the outcome of any test procedure. It is therefore clear that there can be two kinds of errors of judgement. First, one can reject the null hypothesis when it is true. Second, one can fail to reject the null hypothesis when it is false. These two kinds of error are called Type I and Type II errors and are defined as follows:

when it is true. A Type II error occurs if the null hypothesis is not rejected (or accepted) when it is false.

The four possibilities are described in the table below:

Table I

Two possible outcomes for H_0	Decision taken	
	Do Not reject H_0	Reject H_0
H_0 is true	Correct decision	Type I error
H_0 is not true	Type II error	Correct decision

To familiarise you more with the notions of Type I and Type II error, let us look at the problem of 'Milk packets' we discussed in the Unit 5. Suppose we state the hypothesis H_0 and H_1 as

H_0 : The machine is set for 1 litre. (i.e. the average quantity of milk is 1 litre)

H_1 : The average is not 1 litre.

Let us see what are the four possible situations.

Case-1: Suppose that the hypothesis H_0 is actually (really) true. Then if we decide to accept the hypothesis ' H_0 is true', then we have made the right decision. That means if the machine is really set for 1 litre, and if we accept this following our test procedure, then we have made a correct decision.

Case 2: Suppose that, H_0 is really true and on the basis of our procedure we reject it, then we have made a mistake. That means if the machine is really set for 1 litre and our decision is to reject this and confront the milk man, then we are making a **mistake/error**.

Case 3: Suppose that the hypothesis H_0 is actually false. If we accept such a hypothesis, then we have made a mistake. That means if the machine is not properly set for 1 litre, and based on the procedure we conclude that it is set for 1 litre, then we commit a **mistake/error**.

Case 4: Suppose that the H_0 is actually false. If we reject H_0 then our conclusion is correct. That means if the machine is not properly set, and our procedure also concludes so, then we would be fully justified in demanding an explanation from the milk man. Thus there are two situations in which an error occurs.

1) **Hypothesis is true, but we reject it. This is called Type 1 error.**

2) **Hypothesis is false, but we accept it. This is called Type 2 error.**

Now you can try some exercises to see how much you have followed.

E1) Suppose you have a cough. You open the medicine cupboard and find an unmarked bottle. You have a hunch that it is not a cough medicine, rather, it is some poison. If you are using hypothesis test to arrive at a decision, how will you state your null hypothesis and alternative hypothesis? What are the two situations that lead to Type 1 and Type II errors?

E2) If you test a hypothesis and reject the null hypothesis in favour of the alternative hypothesis, does your test prove that the alternative hypothesis is correct? Justify your answer.

Now that we are aware of the types of errors, our aim is to reduce the probability of their occurrence. It is, of course, not possible to eliminate both the errors at the same time.

Depending on the problem in hand, we'll have to choose the type of error which we may prefer to have. For this, we have to look at the consequences: We may have a

Depending on the problem in hand, we'll have to choose the type of error which we may prefer to have. For this, we have to look at the consequences: We may have a situation where, if we make a Type 1 error, we lose Rs.100, and if we make a Type 2 error, then we lose Rs.10,000. In this case we'll try to eliminate the Type 2 error, since it's more expensive.

Similarly let us look at the situation in E1. You open the medicine cupboard and find an unmarked bottle. You have a hunch that it is not a cough medicine, rather it's some poison! You are not sure though. If you reject this hunch and take it to be the cough medicine, when actually it is a poison, then you have made a Type 1 error. Of course, being dead, you would be beyond caring about errors by then! On the other hand, if it is really the cough medicine, but you accept your hunch and refuse to drink it, the only consequence will be that your cough would last a little longer. In this case you would surely opt for the Type 2 error.

Often we have to properly balance the two types of error.

We denote by α , the probability of committing an error of Type 1, and by β , the probability of an error of Type 2.

Let us look at another situation suppose our supplier has sent us 5000 pens. We want to decide whether to accept or reject this lot. We would like all the pens to be in working order, but also realise that there would be some defective ones. We are ready to allow 5% defective. It is not possible to test all the pens. So we decide to test a sample of 20. We now have to decide the criterion of acceptance or rejection of the lot. For example, we may decide to accept the lot if we find at most two defectives in the sample. So we reject the lot (equivalently, decide that the lot has more than 5 % defectives) if we find 2 or more defectives in the sample. If the hypothesis that there are 5 % defectives were true, we can find the probability of rejecting the lot, that is, $P(2 \text{ or more pens are defective})$ by using a binomial probability distribution with $n = 20$, $p = 0.05$. This probability is α , the probability of committing an error of Type 1. The computation of β is not always possible. Because as in this example if $p \neq 0.05$, there are infinitely many alternatives for p . Now, in this example, as soon as we fix the criterion for rejection, α is fixed.

Thus, before start testing a hypothesis, we fix or specify the value of α , or equivalently the critical region. α is called the **level of significance** of the test.

So, it is in our hands to keep the probability of Type 1 error low.

So far we have discussed that, to carry out a test for making decision, we have to choose between the null hypothesis and the alternative. We also note that we have to be careful in making decision because it can lead to errors like Type 1 and Type 2 discussed earlier. So, the tests should be such that it should be possible to measure these errors and to some extent, at least be reduced.

Let us now go back to the problem of the company manufacturing aluminium rods. Let us see how the statistician conducts a test. Note that in this case the statistician actually wants to test whether there is any change in the mean diameter of the rods. Essentially he is testing the change in the parameter mean.

Note that the change can occur in both directions. Either the mean diameter can be very large or it can very small, and neither is desirable. This is why we formulated a two sided composite alternative.

Let us now see how the statistician proceeds. The first step is to state H_0 and H_1 which has been done earlier. We restate it as

$$H_0 : \mu = 2, \quad H_1 : \mu \neq 2$$

where μ is the population mean.

Since the random sample of observations $X_1, X_2 \dots X_n$ are assumed to be i.i.d, with the common distribution function as $\text{Normal}(\mu, \sigma)$, the distribution of \bar{X} is normal and if the null hypothesis is true (That is, if the population mean is 2cm.), the sampling distribution of the sample mean is as shown in Fig.1. We know from Unit 4 that the sampling distribution of mean is normally distributed and has a mean of 2 cm. and a standard deviation $\frac{\sigma}{\sqrt{n}}$ (where σ is the population standard deviation and n is the sample size.) How does this information help the company to decide on whether its hypothesis is acceptable or not? First, they have to choose a level of significance which is "reasonable". Let us say this is $\alpha = 0.05$ i.e. 5%. This α defines a region, which is the total shaded region in the figure below (See Fig.1 below). The unshaded portion in the figure below is called the critical region. Let us see how this region is calculated. Take $\frac{1 - \alpha}{2} = \frac{0.95}{2} = 0.4750$. Look at the normal distribution table, and see where the value 0.4750 is given. You will find this listed in the row 1.9 and column 0.06. Thus it corresponds to $z = 1.96$. So, we consider the region below the curve, and bounded by $\mu_0 \pm 1.96\sigma/\sqrt{n}$ where μ_0 is the hypothetical mean (Here $\mu_0 = 2$). This region is the critical region shown in Fig. 1. The values $z = 1.96$ and $z = -1.96$ are called the **critical values**.

Now we calculate the sample mean and check whether this mean lies within the critical region or outside.

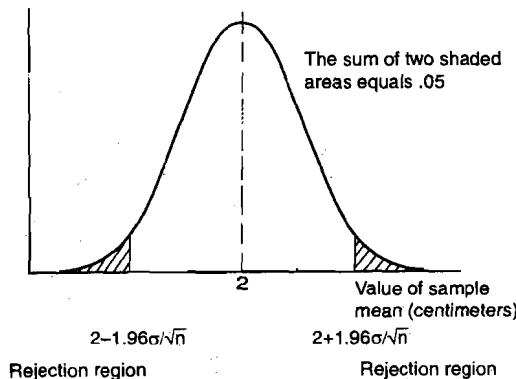


Fig.1

Critical region is calculated using the confidence level discussed in Unit 5. If it lies outside the critical region, then the decision is to reject H_0 in favour of H_1 . That means the assumption that the mean diameter is 2 cm. is rejected. If it lies inside the critical region, then the conclusion is that we cannot reject H_0 . That is, there is insufficient evidence to conclude that the population μ is not 2cm. Even if our sample statistic Fig.1 does fall in the unshaded region (the region that makes up 95% of the area under the curve), this does not prove that our null hypothesis (H_0) is true; it simply does not provide statistical evidence to reject it, why? Because the only way in which the hypothesis can be accepted with certainty is for us to know the population parameter, unfortunately this is not possible. Therefore, whenever we say that we accept null hypothesis, we actually mean that there is not sufficient statistical evidence to reject it.

Let us now summarise the important steps that we have discussed in the testing procedure of any hypothesis.

- 1) Formulate the null and alternate hypothesis .
- 2) Specify the significance level of the test (i.e. fix α).
- 3) Choose a test statistic
- 4) Find the critical region
- 5) Collect the sample and calculate the numerical value of the test statistic based on the sample of observations.

6) Conclusion:

- i) If the numerical value of the test statistic falls in the rejection region, we reject the null hypothesis and conclude that the alternative hypothesis is true. We know that the hypothesis-testing process will lead to this conclusion incorrectly (Type I error) only $100\alpha\%$ of the time when H_0 is true.
- ii) If the test statistic does not fall in the rejection region, we do not reject H_0 . Thus, we reserve judgement about which hypothesis is true. We do not conclude that the null hypothesis is true, because we do not (in general) know the probability β that our test procedure will lead to an incorrect acceptance of H_0 (Type II error).

In the next section we shall illustrate these steps with many examples.

Let us look at another example.

Suppose a light bulb manufacturer believes and advertises that her bulbs have a mean life-time of 1500 hours, and she wants to test her belief. So She formulates the null hypothesis as $H_0 : \mu = 1500$ hours (h)and the alternative as $H_1 : \mu \neq 1500$ h. However, she realises that her customers won't complain if the life of the bulbs they buy from her have a life span of more than 1500 hrs. So she reformulates her problem as

Test $H_0 : \mu = 1500$ h against $H_1 : \mu < 1500$ h.

So in this case we reject H_0 only if the mean life of the sampled bulbs is significantly below 1,000 hours. This situation is illustrated in Fig. 2.

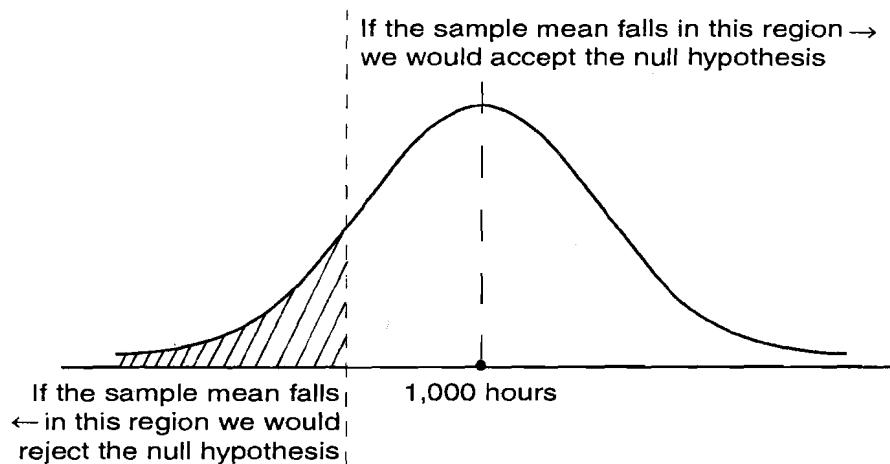


Fig. 2

Note that in this situation, the rejection region is in the left tail of the distribution of the sample mean, and so we call this test a left-tailed (or one-tailed) test.

A left-tailed test is one of the two kinds of one-tailed test. As you have probably guessed, the other kind of one-tailed test is right-tailed test which is used in the situations as given below:

A sales manager has asked her sales people to observe a limit on travelling expenses. The manager hopes to keep expenses to an average of Rs. 100 per salesperson per day. One month after the limit is imposed, a sample of submitted daily expenses is taken to see whether the limit is being observed. Here the null hypothesis is $H_0 = \mu = \text{Rs.}100$, but the manager is concerned only with excessively high expenses. Thus, the appropriate alternative hypothesis in this case is $H_1 : \mu > \text{Rs.}100$. So, in this case the null hypothesis is rejected (and corrective measures taken) only if the sample mean is significantly higher than Rs. 100. The situation is illustrated in Fig. 3.(See next page.) Note that in this situation, the rejection region is in the right-tail of the distribution of the sample means, and so we call this test a right-tailed test.

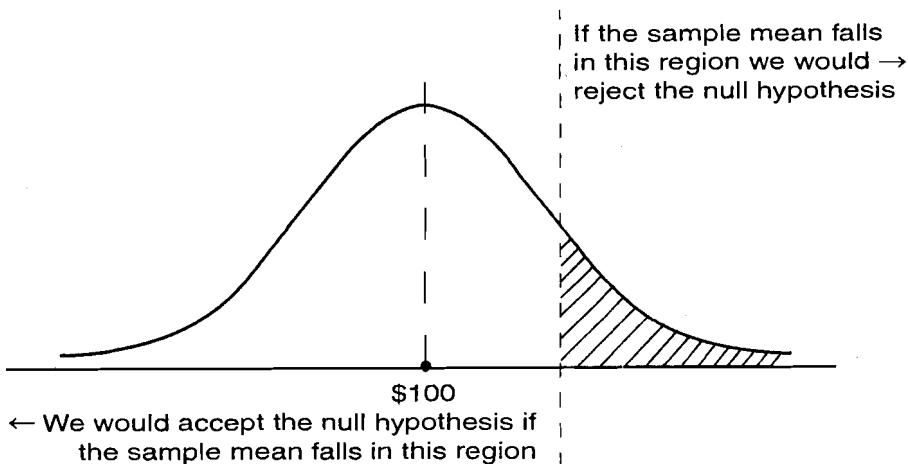


Fig. 3

Let us now summarise our discussion on the different types of tests.

The first case, where we test $H_0 : \mu = 1500\text{h}$ against $H_1 : \mu \neq 1500\text{h}$, we need to use a **two tailed test**; in the second case where we test $H_0 : \mu = 1500\text{h}$ against $H_1 : \mu < 1500\text{h}$, we use a **one-tailed test** (i.e. left-tailed test), and in the third case where we test $\mu = \text{Rs. } 100.00$ against $H_1 : \mu > \text{Rs. } 100.00$, we use a one-tailed test (i.e. a right-tailed test.).

So, in a 2-tailed test we are concerned about any difference from the hypothetical value of the parameter, whereas in a 1-tail test we are concerned with values which are either lower or higher than the hypothetical value. The difference would be clear through the examples. Try this exercise now.

- E3) Radhika, a highway safety engineer, decides to test the load-bearing capacity of a bridge that is 20 years old. Considerable data are available from similar tests on the same type of bridge. Which is appropriate, a one-tailed or two-tailed test? If the minimum load-bearing capacity of the bridge must be 10 tons, what are the null and alternate hypothesis?

In the next section we shall now show how hypotheses about parameters are tested through some examples. In this unit we shall deal with only those cases, where the population is taken to be normally distributed.

We start with tests involving the parameter mean of the population.

6.3 TESTS ABOUT THE MEAN (z-test and t-test)

In this section we show how to test whether a sample is drawn from a population with a given mean. We'll also show how to test whether two samples belong to the same population or not.

6.3.1 Comparing the Mean of a Population and a Sample

We first illustrate the technique when the variance of the population is known.

Let us look at a problem.

Problem 1: The mean marks obtained by the students of a mathematics course in IGNOU is 54.5 with a standard deviation 8.0. At one of the study centres, where 100

students took the examination, the mean marks were 55.9. Are the students of this study centre significantly 1) different from 2) better than, the rest of the students of that course in IGNOU at 0.01 level?

Solution: We shall first solve (1). Let X_1, X_2, \dots, X_{100} denote the random variables which are the marks obtained by the 100 students. We assume that these random variables are i.i.d as normal with mean μ and standard deviation $\sigma_0 = 8.0$. We want to test if $\mu = 54.5$ against the alternative that $\mu \neq 54.5$.

Let us apply the six steps given in the earlier section.

- 1) Here $H_0 : \mu = 54.5$, $H_1 : \mu \neq 54.5$ and $\alpha = 0.01$. It is a two-tailed test. The test statistic is \bar{X} . Now we have to fix the critical region as discussed in the previous section. You know that if you take samples of size 100, then the sample means \bar{X} are normally distributed with mean $\mu = 54.5$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{100}} = 0.8. \text{ This means that } \frac{\bar{X} - 54.5}{0.8} \text{ follows a standard normal}$$

distribution. Now, since $\alpha = 0.01$ (1 % level), $\frac{1-\alpha}{2} = 0.4950$. So we find the z-value corresponding to 0.4950, which we denote by $z_{0.4950}$, such that

$$P\left[-z_{0.4950} < \frac{\bar{X} - 54.5}{0.8} < z_{0.4950}\right] = 1 - \alpha = 0.99.$$

Now we look at the normal distribution table given at the end of the block and see where the value 0.4950 is given. This is listed in the row for 2.5 and column for 0.08. That is $z_{0.4950} = 2.58$.

Therefore, we have $P\left[-2.58 < Z = \frac{\bar{X} - 54.5}{0.8} < 2.58\right] = 0.99$. Also 2.58 and -2.58 are the critical values and the accepted region is as shown in Fig. 4.

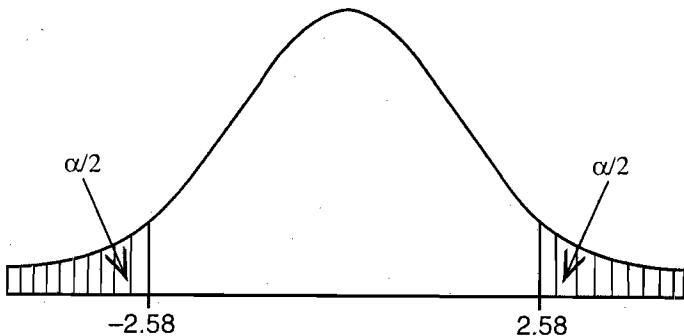


Fig.4: The total area of rejection (the shaded part) should be α .
It is already given in the problem that the value of the sample statistic \bar{X} is 55.9. The corresponding z-value is

$$\begin{aligned} z &= \frac{55.9 - 54.5}{0.8} \\ &= 1.75 \end{aligned}$$

Since $z = 1.75$ is in the acceptance region, we accept the hypothesis, or more precisely, we fail to reject it.

This means the difference in the two mean marks is not significant enough to suggest that the students at that study centre are different from the rest of the IGNOU students.

- 2) Here we apply a one-tailed test.

$$H_0 : \mu = 54.5, H_1 : \mu > 54.5$$

Since we are interested in knowing whether the group of students is **better**, we need to look at only those values of \bar{X} which are higher than 54.5. So we find the appropriate z-value such that the rejection area is the shaded portion α as shown Fig. 5 and the acceptance region is the unshaded portion. From the normal distribution table, we determine that the value of z for 40% of the area under the

curve is $z = 2.33$. Here we have only one critical value 2.33. Also we have

$$P \left[Z = \frac{\bar{X} - 54.5}{0.8} < 2.33 \right] = 0.99.$$

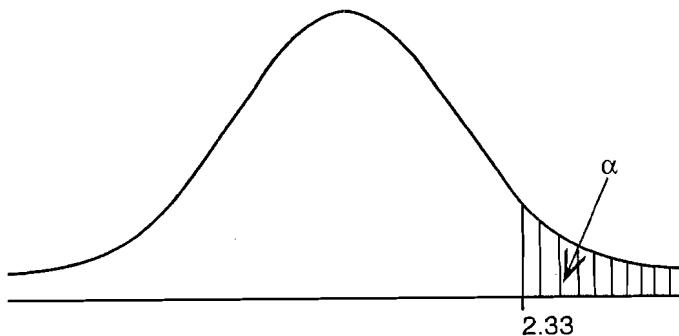


Fig.5

The shaded region is given in Fig.5. Again, the test statistic here is 1.75, which is less than 2.33 and therefore falls in the acceptance region and so we fail to reject H_0 . That is, we conclude that the students from that study centre are not better than other students at 1 % level of significance.

————— X —————

You can see that the basic technique is to come up with an acceptance region by using the given level of significance (α), after noting whether it is a 1-tailed or a 2-tailed test. We then accept or reject the null hypothesis, depending on whether the given value of \bar{x} falls in the region or not. You must have noticed here the similarity with the computation of confidence intervals discussed in Unit 5. There we find an interval around \bar{x} . Here we find an interval around μ_1 , the hypothetical mean.

In the above example, α was given to be 0.01. If $\alpha = 0.05$, the only difference is in the value of z that we read from the tables. This value for a 2-tailed test is 1.96 and for a 1-tailed test is 1.64. You have been using these same values of z for calculating 95% and 99% confidence intervals in Unit 5.

In the following table, we give the critical values for both one-tailed and two-tailed tests at three significance levels, $\alpha = 0.1, 0.01$ and 0.05 .

Table 2

level of significance α	0.1 (90%)	0.05 (95%)	0.01 (99%)
critical value for left-tailed test	-1.28	-1.64	-2.33
critical value for right-tailed test	1.28	1.64	2.33
critical values for two-tailed tests	-1.64 and 1.64	-1.96 and 1.96	-2.58 and 2.58

Note: - The test procedure applied in problems 1 and 2 is called **z-test**. Note that z-test is applicable when we assume that the random variables representing the sample of observations, X_1, X_2, \dots, X_n are i.i.d normal(μ, σ^2) where σ^2 is assumed to be known. It is also valid if the sample size n is large i.e. $n > 30$ (by the central limit theorem) even if the common distribution is not normal or if σ^2 is unknown and has to be estimated from the sample.

Let us consider the following problem.

Problem 2: The manufacturer of an antacid claims that it relieves discomfort in 5 minutes (with a standard deviation of 2 minutes). Ten people volunteer to take it to test

the claim. The average time to get relief was 7.5 minutes. Do you accept the claim at a 10% level?

Solution: We shall assume that the time to get relief is distributed normally with unknown mean μ and known standard deviation 2. Here $H_0 : \mu = 5$ mins, $H_1 : \mu \neq 5$ mins.

Here the test is two-tailed and from Table 1 we get that the critical values are 1.64 and -1.64. Also we have

$$P\left[-1.64 < \frac{\bar{X} - 5}{2/\sqrt{10}} < 1.64\right] = 0.9$$

Since $\frac{\bar{X} - 5}{2/\sqrt{10}} = \frac{7.5 - 5}{2/\sqrt{10}} = 3.95$ is greater than 1.64, does not lie in the acceptance region, so we reject H_0 , and conclude that the claim is not justified at 10% level.

————— X —————

In the next problem we consider the situation where the sample size n is large i.e. $n \geq 30$ and σ is unknown whereas the sample standard deviation s is given. Note that in this situation we apply z-test by replacing σ by s .

Let us see an example.

Problem 3: A consumer magazine, when comparing various brands of paints, stated that the drying time of one particular brand was found to be four hours. The manufacturer was not particularly pleased with this and consequently modified the paint to try to reduce the drying time. The paint was then tested by a random sample of 40 customers all of whom were decorating their living rooms. For this sample the mean drying time in hours was found to be 3.85 and the sample standard deviation was 0.55.

- a) Analyse the sample data using the one-sided z-test.
- b) Find a 95% confidence interval for the population mean of the drying times for the modified paint.
- c) What can you conclude about the drying time of the modified paint?

Solution:

- a) We want to test the hypothesis that the population mean (μ) of the drying times (in hours) of the modified paint is equal to 4. The appropriate null hypothesis is therefore

$$H_0 : \mu = 4$$

Since the manufacturer is looking for a reduction in the drying time (at least he does not expect that it should have increased!), the alternative hypothesis should be one-sided, giving

$$H_1 : \mu < 4.$$

The test statistic is

$$z = \frac{\bar{x} - 4}{s}$$

Here we do not know the population variance. But we know that the sample standard deviation is 0.55. Therefore using the estimation, we can find an estimate of the population S.D as $\sigma = \frac{s}{\sqrt{n}} = \frac{0.55}{\sqrt{40}} = 0.0869626$.

Note that here the test is left-sided and $\alpha = 0.05$. From Table 1, we get the critical value as -1.64.

Since the sample mean \bar{x} is 3.85, we get that $z = \frac{3.85 - 4}{0.869626} \sim -1.72$ (cutting to two decimal places).

Since the value of the test statistic is -1.72 is less than the critical value -1.64 , we reject H_0 in favour of H_1 at the 5% significance level and conclude that it looks as though the population mean of the drying times for the modified paint is less than four hours as the manufacturer hoped. Of course, this result only applies to paint drying in living rooms.

- b) A 95% confidence interval for μ is given by $(\bar{x} - 1.96 s/\sqrt{n}, \bar{x} + 1.96 s/\sqrt{n})$ which is

$$[3.85 - (1.96 \times 0.0869626), 3.85 + (1.96 \times 0.0869626)]$$

giving

$$[3.67, 4.02]$$

c)



Fig.6

As in the case of the earlier problem, it is always a good idea to follow up a hypothesis by finding a confidence interval since such an estimate provides useful extra information. Of course, with some hypothesis tests you may not be able to do this with one reason or other.

You can try some exercises now.

- E4) The breaking strengths of cables made by a company had a mean of 1800 N. The company then adopted a new technique which is believed to increase the breaking strengths. 50 cables made by the new technique were tested to see if the belief is justified. or not. The mean breaking strength of these 50 is found to be 1850N with a standard deviation of 100N. Is the belief justified at a) 5% level b) 1% level.
- E5) As part of a survey on drivers' reaction times for a driving magazine, 300 drivers were subjected to the following test : each driver was asked to press a lever with his/her foot in response to a flashing light. The reaction times (in seconds) were recorded and the sample mean was found to be 0.83. The sample standard deviation was 0.31. What can you conclude about drivers' reaction times?

So far we have seen that z-test can be used when σ is known whether the sample size is large or small. But when the sample size is small, z-test can not be applied when σ is not known and in the case we use a test based on t-distribution instead of normal distribution. In the following problem we illustrate the use of t-distribution.

Problem 4: A machine manufactures standard weights to be used in weighing scales. To check if the machine is working properly, a random sample of five 2-kg. weights was taken. Each 2-kg. weight was weighed on a special scale and the actual weights were found to have a mean of 1.962 kg. and a standard deviation of 0.038 kg. If $\alpha = 0.05$, can you say that the machine is in proper working order?

Solution: Assume that the random observations X_1, X_2, \dots, X_5 are distributed as i.i.d normal (μ, σ^2) Our null hypothesis is that the machine is in proper working order. That

is $H_0 : \mu = 2\text{kg}$, $H_1 : \mu \neq 2\text{kg}$.

Note that here we do not know the population standard deviation (This means that we are testing a composite hypothesis against a two sided composite alternative.)

$$\text{Now we form the test statistic } t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.962 - 2}{0.038/\sqrt{5}} = -2.24$$

This follows a t distribution with 4 d.f. Now this is a 2-tailed test. So we need to get $t_{.975}$ for 4 d.f. from the t-table. We see that $t_{.975} = 2.776$. So the acceptance region is $(-2.776, 2.776)$ and -2.24 falls in this region.

So, we conclude that at 5% level of significance, the mean of the weights manufactured by the machine is 2Kg. That is, we say that the machine is in proper working order. In other words, we say that we have not found sufficient evidence to suggest that the machine is not working properly.

The situation in the above example called for a 2-tailed test because we were interested in testing against a two sided alternative. whether the weights were any different from 2 kg. The next example presents the case for a 1-tailed test.

Problem 5: A management school claims that the starting salaries for its graduates average Rs.10,000 or more per month. A random sample of 7 students who had recently graduated, showed an average salary of Rs.9700 with a standard deviation of Rs.306. At a 5% level of significance would you accept the claim?

Solution: Here you would be interested in checking if the average salaries are less than claimed. (If the average is more than Rs.10,000, the claim is justified of course). So

$$H_0 : \mu = \text{Rs.}10,000, H_1 : \mu < \text{Rs.}10,000$$

$$\text{Here } t = \frac{9700 - 10,000}{306/\sqrt{7}} = -2.59.$$

Since we have to apply a 1-tail test to check the lower end, we have to get a lower limit of t corresponding to 5% level with 6 d.f. Now, to get $t_{0.05}$, we find $t_{.95}$ with 6 d.f. and take its negative. So, $t_{.05} = -1.943$. Since $t = -2.59$ falls in the rejection region or the critical region, we reject H_c . This means that the claim is not justified at 5% level.

See if you can do this exercise now.

- E6) The specifications for the production of a certain alloy call for 23.2% copper. In 10 analyses, the mean copper content was found to be ~~23.5 n of 0.24%~~. Can we conclude that the product meets the specifications if $\alpha = 0.05$?
- E7) The diameters of bolts manufactured by a machine are known to have a ~~standard deviation of 0.0002 cm~~. A random sample of 10 bolts has an average diameter of ~~and standard deviation~~ 0.5046 cm. Test the hypothesis that the true mean diameter of bolts is 0.51 cm, using $\alpha = 0.01$.

In all the examples and exercises till now, we have been using a function of the sample mean to test hypothesis about the population mean. Suppose now we want to compare the effect of two medicines in providing relief from pain. The duration for which a medicine provides relief from pain can be observed and we may assume that the duration for medicine 1 is a random variable X that is distributed as normal with mean μ_X and standard deviation σ_X . Similarly the duration for which medicine 2 provides relief is a random variable Y which is assumed to be distributed as normal with mean

μ_Y and standard deviation σ_Y . We want to test the null hypothesis (the hypothesis of no difference) $H_0 : \mu_X = \mu_Y$ against the alternative that H_0 is not true. We discuss similar problems in the next sub-section.

6.3.2 Difference in the Means of Two Populations

Here we shall be dealing with two populations or two variables, X_1, X_2 assumed to be normally distributed with means μ_1, μ_2 and variances σ_1^2 and σ_2^2 , respectively. We shall have to divide our discussion in two parts: 1) when σ_1, σ_2 are known, and 2) when σ_1, σ_2 are unknown. In Sec.6.3.1 you have seen that the test statistic for the problems discussed there follows a standard normal distribution when the standard deviation is known. You have also noted that it follows the t-distribution in case the standard deviation is not known.

We start with the first case now.

σ_1, σ_2 known:

We shall illustrate the method through an example.

Example 1: Suppose we want to investigate the following:

“Was there a difference in the performance of male and female students of IGNOU in the examination for Mathematics Elective courses in a particular year, say 2000? If so, what was the difference?

Let us see how we can use z-test to find an answer to this question.

We shall first set up the null and alternate hypothesis in terms of the population means as follows:

H_0 : The mean examination mark for the population of all male students is the same as the mean examination mark for the population of all female students.

H_1 : The mean examination mark for the population of all male students is not the same as the mean examination mark for the population of all female students.

We can now introduce some symbols that enable us to express these hypotheses more concisely. We let

μ_m denote the population mean of the marks of all male students

and

μ_f denote the population mean of the marks of all female students.

Then H_0 and H_1 become

$$H_0 : \mu_m = \mu_f$$

$$H_1 : \mu_m \neq \mu_f.$$

We also introduce now some other symbols that will be useful in our analysis of the sample data. Let

σ_m and σ_f denote the population standard deviations of the marks of all male and female students respectively;

\bar{X}_m and \bar{X}_f denote the two sample means;

s_m and s_f denote the two sample standard deviations;

n_m and n_f denote the two sample sizes.

Now suppose we take a sample of 150 male students and 100 female students. Then the table in the next page gives the sample means and sample s.d.'s for the samples taken.

We may use $\frac{\bar{X}_m - \bar{X}_f}{SSD((\bar{X})_m - \bar{X}_f)}$ as the test statistic, where SSD denotes the estimate

$$\sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}.$$

	Sample size	Sample mean	Sample standard deviation
Male	$n_m = 150$	$\bar{x}_m = 57.88$	$s_m = 20.00$
Female	$n_f = 100$	$\bar{x}_f = 63.73$	$s_f = 22.10$

Here, we need to find the distribution of $\frac{\bar{X}_m - \bar{X}_f}{\text{SSD}(\bar{X}_m - \bar{X}_f)}$. In our problem we have assumed that the marks obtained by the male or female students follow a normal distribution. Therefore the distributions of the sample means are also normal with appropriate means and standard deviations. Also $\bar{X}_m - \bar{X}_f$ will have a normal distribution with an appropriate mean and standard deviation. However the distribution of $\frac{\bar{X}_m - \bar{X}_f}{\text{SSD}(\bar{X}_m - \bar{X}_f)}$ is unknown but can be taken approximately as standard normal by using the central limit theorem, provided the sample sizes n_m and n_f are large, say more than 30. From Unit 4 you know that the sampling distribution \bar{X}_m of the mean \bar{X}_m is normal with mean μ_m and standard deviation $SE = \sigma_m / \sqrt{n_m}$. The standard deviation is given by

$$SE = \frac{\sigma_m}{\sqrt{n_m}} = \frac{\sigma_m}{\sqrt{150}}.$$

(See Fig.7)

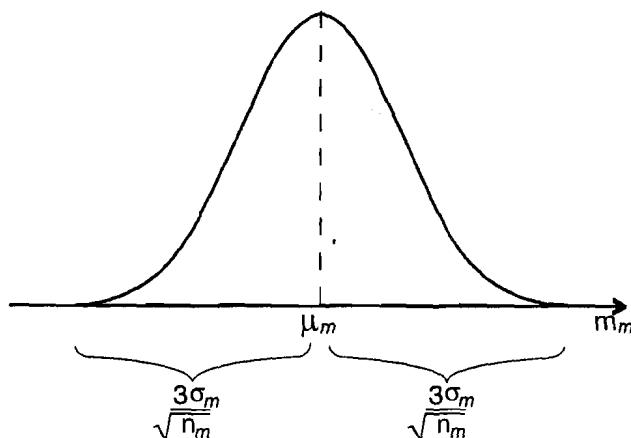


Fig. 7 Sampling distribution of \bar{X}_m .

Similarly the sampling distribution \bar{X}_f is normal with mean μ_f and standard deviation

$$SE = \frac{\sigma_f}{\sqrt{n_f}} = \frac{\sigma_f}{\sqrt{100}}.$$

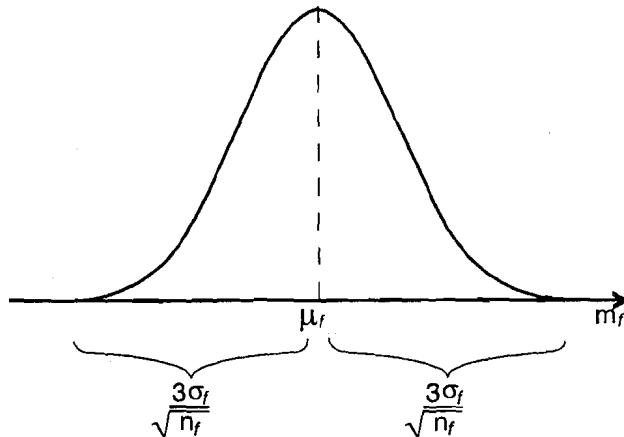


Fig. 8: Sampling distribution of \bar{X}_f

Now, consider all the possible pairs of samples of sizes 150 and 100 respectively that

we could select from the two populations of male and female students. For each of these pairs of samples (and there is a huge number of possible combinations) we can calculate the **difference between the sample means** ($\mu_m - \mu_f$) and by considering all the combinations of such samples we obtain the **sampling distribution of**

$\frac{\bar{X}_m - \bar{X}_f}{\text{SSD}((\bar{X}_m) - (\bar{X}_f))}$. It turns out that the distribution of the statistic is also approximately Normal.

Moreover, the mean of this statistic is

$$\mu_m - \mu_f$$

and its standard deviation is given by

$$\text{SE} = \sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}$$

The sample estimate of the standard deviation is

$$\text{SSD} = \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}$$

and it is this estimate that we shall use here. The sampling distribution of $(\bar{X}_m - \bar{X}_f)$ is given in Fig.9.

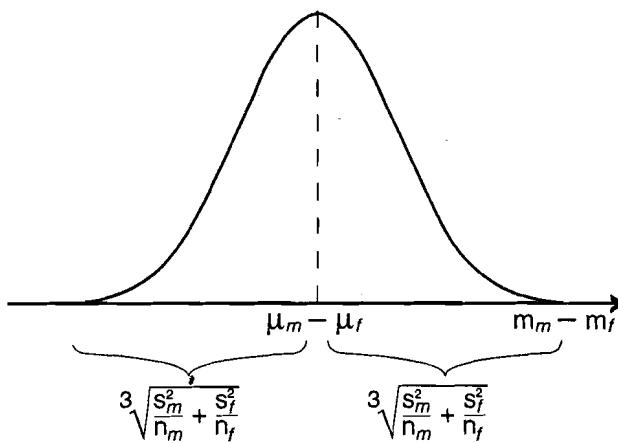


Fig.9/Sampling distribution of $\bar{X}_m - \bar{X}_f$

Before we proceed further we make a note.

Note: If we knew the standard deviations σ_m and σ_f then we didn't have to replace them by the sample estimates. In that case our test statistic will be given by $\frac{(\bar{X}_m - \bar{X}_f) - (\mu_m - \mu_f)}{\sqrt{\frac{\sigma_m^2}{n_m} + \frac{\sigma_f^2}{n_f}}}$

Now we apply the z-test based on the statistic $\bar{X}_m - \bar{X}_f$ to test the given null hypothesis. The hypotheses under consideration are

$$H_0 : \mu_m = \mu_f \text{ and } H_1 : \mu_m \neq \mu_f$$

and, because

$$\mu_m = \mu_f$$

can be rewritten as

$$\mu_m - \mu_f = 0,$$

we can express H_0 and H_1 in terms of the difference between μ_m and μ_f as follows

$$\begin{aligned} H_0 : \mu_m - \mu_f &= 0 \\ H_1 : \mu_m - \mu_f &\neq 0. \end{aligned}$$

Now we just apply the z-test discussed in the earlier section replacing μ by $\mu_m - \mu_f$ and \bar{X} by $\bar{X}_m - \bar{X}_f$.

Hence the test statistic is

$$z = \frac{(\bar{x}_m - \bar{x}_f) - (\mu_m - \mu_f)}{SE}$$

and, since the null hypothesis is $\mu_m - \mu_f = 0$, this simplifies to

$$z = \frac{\bar{x}_m - \bar{x}_f}{SE}$$

$$\text{where, } SE = \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}$$

Now from the data given in Fig. we can calculate the z-value. We have

$$\begin{aligned} SE &= \sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}} \\ &= \sqrt{\frac{(20.00)^2}{150} + \frac{(22.10)^2}{100}} \\ &= \sqrt{2.6666667 + 4.8841} \\ &= \sqrt{7.5507667} \\ &= 2.7478658 \\ \text{and } \bar{x}_m - \bar{x}_f &= 57.88 - 63.73 = -5.85 \\ \text{so } z &= \frac{-5.85}{2.7478659} \sim -2.212 \end{aligned}$$

Hence the test statistic is $z = -2.12$ and now the procedure is exactly the same as it was in previous Section.

Since the test statistic -2.12 is less than the critical value -1.96 , we reject H_0 in favour of H_1 at the 5% significance level and conclude that $\mu_m - \mu_f$ does not seem to be equal to zero. Indeed, because -2.12 is less than -1.96 , it looks as though $\mu_m - \mu_f$ is less than zero and this means that μ_m seems to be less than μ_f . This suggests that there is a difference in examination performance between the sexes; it seems that the females perform better than the males.

* * *

We shall now summarise the steps used in the example above, for conducting the test.

Our aim here is to test whether two given normally distributed random variables have the same mean or not. For this, we first state H_0 and H_1 as

$$H_0 : \mu_1 = \mu_2 \text{ i.e. } (\mu_1 - \mu_2) = 0; \quad H_1 : \mu_1 \neq \mu_2.$$

Then take a random sample of size n_1 from the first population and a random sample of size n_2 from the second. We find the means of these samples: \bar{X}_1 and \bar{X}_2 . Here The

difference $\bar{X}_1 - \bar{X}_2$ is normally distributed with mean $\mu_1 - \mu_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

So, if H_0 is true, then the test statistic $Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ is distributed as standard

normal.

So we find $Z_{1-\alpha/2}$ where α is the level of significance. We then reject H_0 if $Z > Z_{1-\alpha/2}$, or $Z < -Z_{1-\alpha/2}$.

If we have a situation where a one-tailed test is to be applied, say,

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 < \mu_2,$$

then we find $Z_{1-\alpha}$ and reject H_0 (that is, accept H_1) if $Z < -Z_{1-\alpha}$.

Study the following examples carefully now, so that you can solve the exercises which come later.

Problem 6: Two machines are used to fill cans with 200 ml. of a drink. The filling processes are assumed to be normal, with standard deviations $\sigma_1 = 0.2$ and $\sigma_2 = 0.25$. The quality control department wants to check if the two machines fill the same volume. A random sample is taken from the output of each machine:

M ₁	M ₂
200.3 200.1 200.4 198.9 200.5	200.2 200.3 199.7 200.4 199.4
199.2 200.5 200.2 200.2 199.5	200.2 200.1 200.1 199.8 200

what is your conclusion? Use $\alpha = 0.05$.

Solution: Note that the standard deviations are assumed to be known. We state the hypothesis as

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_A : \mu_1 &\neq \mu_2 \end{aligned}$$

The test statistics is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

From the data given in the two samples we have

$$\bar{x}_1 = 199.98 \text{ and } \bar{x}_2 = 200.02$$

$$z = \frac{199.98 - 200.02}{\sqrt{\frac{0.2^2}{10} + \frac{0.25^2}{10}}} = 0.395$$

Now, for $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$. Since $z < z_{1-\alpha/2}$, we accept H_0 .

That is, we have not found any evidence to suggest that the two machines fill different volumes of the drink.

————— X —————

Problem 7: Two different formulations of petrol are tested to study their road octane numbers. The variance for Formulation 1 is $\sigma_1^2 = 1.8$ and for Formulation 2 is $\sigma_2^2 = 1.2$. Two random samples of size $n_1 = 15$ and $n_2 = 20$ are taken. The mean road octane numbers are $\bar{x}_1 = 89.6$ and $\bar{x}_2 = 92.5$. Can you say that Formulation 2 produces a higher road octane number than Formulation 1 if 1) $\alpha = 0.05$, 2) $\alpha = 0.01$?

Solution: Here $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$.

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{89.6 - 92.5}{\sqrt{\frac{1.8}{15} + \frac{1.2}{20}}} = -6.84$$

- 1) For $\alpha = 0.05$, $z < -z_{0.95} = -1.64$. Therefore we reject H_0 , that is accept H_1 .

2) For $\alpha = 0.1$, $z < -z_{.99} = -2.33$. Therefore we again reject H_0 , and accept H_1 .

So, at both the levels we say that Formulation 2 produces a higher road octane number than Formulation 1.

— X —

If you have followed these examples, you would certainly be able to these exercises.

E8) A new teaching technique is to be tested. A group of 22 students were taught in the traditional way. Another group of 18 students was taught with the help of the new technique. The two groups were then given a standardised test which is known to have a standard deviation of 25. The mean score of the traditional group was 127 and that of the experimental group was 136. If $\alpha = 0.1$, do you think that the new technique is significantly better?

E9) A psychologist gave a test to decide if male students are as smart as female students. The sample of 40 female students had a mean score of 131 and the sample of 36 males had a mean score of 126. The test has a standard deviation of 16. Is there a difference at 0.01 level of significance?

We have been considering cases where σ_1 and σ_2 are known. If they are not known, they have to be estimated from the sample. If the samples are large, then these estimates are quite close to the real values and so we can use them in forming the test statistic Z. In the next exercise you see one such situation.

E10) A sample of 100 electric light bulbs produced by manufacturer A showed a mean life-time of 1190h and a standard of 90h. A sample of 75 bulbs produced by manufacturer B showed a mean life-time of 1230h and a standard deviation of 120h. a) Is there a difference between the two brands of bulbs at a significance level of 0.05? b) Are the bulbs of manufacturer B superior to those of manufacturer A at the same level?

Now we come to the second case. Here we shall see how to test for the difference between means when the population variance is not known and the sample sizes are small.

σ_1, σ_2 unknown

Before we go any further, we must make an additional assumption. The population variances are unknown here, but whatever they are, we are going to assume that they are equal. This is because, the situation becomes very complicated if the population variances are not equal, and in this course we are not yet ready to tackle it. The common but unknown population variance has to be then estimated from the samples. So, obviously, we have to compute s_1^2 and s_2^2 which are unbiased estimates of σ^2 . Now, though we have assumed the populations variances to be equal, s_1^2 and s_2^2 may not be equal. Therefore, we use them to form a pooled variance, which we then take as a single estimate of σ^2 .

We form $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$, and then compute the test statistic

$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, which is distributed as student's t with $n_1 + n_2 - 2$ degrees of

freedom. Here is an example to illustrate this procedure.

Problem 8: Suppose we have to choose between two types of paving surfaces to be used on a highway. One of the considerations is the stopping distance of cars. The

distance taken by cars travelling at 80 km. per hour to come to a complete stop was measured. The results (in meters) were as follows:

Surface A : $n_1 = 8, \bar{x}_1 = 42.3, s_1^2 = 38.8$

Surface B : $n_2 = 8, \bar{x}_2 = 43.2, s_2^2 = 51.$

Is there a difference in the stopping distance between the two surfaces at 0.05 level of significance?

Solution: $H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2$

$$\text{Now, } s_p^2 = \frac{7(38.8 + 51)}{14} = 44.9$$

$$\text{Then } t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{42.3 - 43.2}{\sqrt{44.9} \sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.25$$

The degrees of freedom are $8 + 8 - 2 = 14$. This is a two-tailed test. So, $t_{.975}$ for 14 d.f. is 2.14. Since t is within the limits, -2.14 and 2.14, we fail to reject H_0 . Therefore, we conclude that at the given level, there is no difference between the two types of surfaces.

————— X —————

We have been testing for the equality of the means of two normally distributed populations by using the sample means. The samples that we had considered were independent samples. That is, the individuals in one sample were not matched or paired with those in the other in any way.

We now look at some situations where the individuals in the samples are paired or the samples are dependent.

For example, suppose the students of a particular class are given a crash course in speed reading. To test whether their reading speed has really increased or not, we need to take a sample and note the reading speeds of the students in this sample both before and after the course. In this case, the two data sets (comprising the before and after reading speeds) are not independent since each observation in the before set is matched with an observation in the after set. Here we compare the score of each student before he/she took the course to his/her score after the course, and try to find out if there is a pattern. If 60% of the students show an increased speed, is it reasonable to call the course a success? We can answer such questions with the help of t-distribution.

We start by calculating the difference for each of the pairs. In this particular example? It means that we take the difference between the reading speeds of each student. We then assume that these values (differences) are normally distributed. Next, we calculate the mean and standard deviation from the sample and apply the t-test. See the following situation.

Problem 9: The following table gives the data on the speeds of 10 students:

Table 3

Before	9.4	10.3	8.4	6.8	7.8	9.8	9.2	11.2	9.4	9.0
After	9.3	10.6	8.8	7.0	7.7	10.0	9.8	11.7	9.7	9.0
Difference	-0.1	0.3	0.4	0.2	-0.1	0.2	0.6	0.5	0.3	0.0

Can you say that the reading course is a success at 0.05 level of significance?

We denote the mean difference in the speeds of the population by \bar{D} and that of the sample by \bar{d} .

Solution: Suppose that $H_0 : \bar{D} = 0$ (i.e. there is not difference in the reading speeds

before and after). $H_1 : \bar{D} > 0$.

Here $\bar{d} = 0.23$, and the standard deviation for the differences (S_d) is 0.241. $n = 10$. The test statistic is:

$$\begin{aligned} t &= \frac{\bar{d}}{S_d/\sqrt{n}} \\ &= \frac{0.23}{0.241/\sqrt{10}} = 3.02. \end{aligned}$$

This is a one-tailed test. Therefore $t_{0.05}$ for 9 degrees of freedom is 1.83. Since $t=3.02$ is more than this upper limit, we reject the null hypothesis and accept the alternative one. So, we can conclude that the speed reading course is helpful in increasing the reading speeds of students.



We are sure you can do these exercises now.

E11) We want to test the effect of a new fertiliser on wheat production. For this, 24 plots of land of equal area were chosen. Half of these were treated with the new fertiliser and the other half were treated with old one. With the new fertiliser, the mean yield was 48 kg. with a standard deviation of 4 kg. With the old fertiliser, the mean yield was 51 kg, with a standard deviation of 3.6 kg. Can we say at 5 % level of significance that there is an improvement in the yield because of the new fertiliser? What will be your conclusion at 1 % level?

E12) A botanist was interested in knowing if there was a difference in the time fruits matured on different parts of a plant, and recorded the day of the first fruit on the top and on the bottom for 15 plants. All the fruits came out during the same month.

Top	3	6	7	5	8	9	10	10	7	8	6	9	10	12	4
Bottom	7	9	5	8	8	10	11	12	6	9	7	13	8	13	8

Is there a significant difference in the time to mature at the 1% significance level?

E13) The pulse rates of 12 people were recorded before and after taking a new drug.

Before	68	71	84	93	67	74	82	77	71	83	62	66
After	71	70	81	97	73	80	90	76	80	79	80	67

Using 10% level, can you say that there is a significant increase in the pulse rate?

When you have solved E11, you might have found that you reject the hypothesis at 5 % level, but accept it at 1 % level. In such a case we say that the results are probably significant, that is, the new fertiliser is probably better than the old one. But we need to get more evidence to make a decision. After these numerous tests about the mean we now turn our attention to the variance of a population.

6.4 TEST ABOUT THE VARIANCE

You must have seen in Unit 4 that if the population is normally distributed with mean μ and variance σ^2 , then the ratio $\frac{(n-1)s^2}{\sigma^2}$ is a χ^2 variable with $(n-1)$ d.f. Now here we are again going to use the χ^2 distribution to test if a given sample could have come from a population with a given variance, σ^2 . Of course, before applying this test, we should make sure that the population is normal. Actually, even when we discussed the

tests about the mean of a population, we had assumed the population to be normal. But we have to be especially careful in this case, because this test is particularly sensitive to the shape of the distribution. If we apply the test to the variance of a non-normal population, then chances are that we may be committing mistakes much more frequently than the α value indicates.

Suppose we have a sample of size n from a normal population. For this test our null hypothesis is that it comes from a population with a given σ^2 . We form the test statistic, $\frac{(n - 1)s^2}{\sigma^2}$. Now we compare the value of the test statistic with χ_{α}^2 for a significance level of α . If the value of the test statistic is greater than or equal to χ_{α}^2 , then we reject H_0 . Otherwise we accept it.

Problem 10: A machine is used to fill 2-kg packages of rice. It is known to have a standard deviation of 12.5 g. To check if the machine has become more erratic, a sample of 20 packages was taken. This showed a standard deviation of 16 g. Is the increase in variability significant at 1) 0.05 and 2) 0.01 levels?

Solution: $H_0 : \sigma = 12.5$ g., $H_1 : \sigma > 12.5$ g.

$$\text{Now, } \chi^2 = \frac{(n - 1)s^2}{\sigma^2} = \frac{19(16^2)}{(12.5)^2} = 31.13$$

Here we have to use a one-tailed test.

1) $\chi_{0.05}^2$ for 19 d.f. is 30.1. So, we reject H_0 at 0.05 level.

2) $\chi_{0.01}^2$ for 19 d.f. is 36.2. So, we fail to reject H_0 at 0.01 level.

Because of the nature of conclusions in 1) and 2), we say that the variability of the machine has probability increased, and recommend that the machine should be examined. (Also see Ex 4)

We now ask you to do a few exercises.

E14) It has been found that the variability of driving speeds among drivers (and not speeding) is the main cause of accidents. It has also been found that the optimum standard deviation of speeds on a highway is 3 km. per hour. A sample of speeds of 16 cars was taken and its s was found to be 14. Is this variability greater than the optimum at 5 % level of significance?

E15) The quality of printing paper depends on the variation in thickness. According to the specifications of a printer, the optimum variance of thickness is 0.0022 cm. A sample of 13 readings of the thickness of the paper supplied showed a S^2 of 0.0031 cm. Should the printer reject the paper at 1 % level of significance?

We shall take up the last category of tests that we are going to discuss here: tests about proportions.

6.5 TESTS ABOUT THE POPULATION PROPORTION

In many applications we come across a binomial variable. These are the situations in which the population is divided in exactly two categories: male/female, good/bad, acceptable/defective, and so on. Here we are interested in the proportion of one of the categories. Again, to estimate the proportion in the population we take the help of a random sample. You have already seen this in Unit 5. Even though the variable is binomial, p is still found to follow a normal distribution. This is of course true, when the sample size is large and p is not too close to either 0 or 1. We use this fact now to

test the hypotheses about the population proportion. You must have realised by now that tests of hypotheses run exactly parallel to the computation of confidence intervals. Before we illustrate the procedure through some examples, here are a few points to remember.

- 1) If π is the proportion of a certain category in the population, and p is that calculated from the sample, then under the above restrictions, $z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$ is a standard normal variable.
- 2) If two samples of sizes n_1 and n_2 are drawn from the same population, and if p_1 and p_2 are the estimates of π obtained from them, then $z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ is a standard normal variable, where $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Let us consider this problem.

Problem 11: A private gallery purchased a rare painting, expecting that it would attract 75 % of its visitors. To verify this, a sample of 60 people was taken. It was found that out of these, 35 had looked at the painting. Do you think that the expectation was justified at 5% level?

Solution: $H_0 : \pi = 0.75, H_1 : \pi \neq 0.75$

$$\text{Here } n = 60 \text{ and } p = \frac{35}{60} = 0.58$$

$$z = \frac{0.58 - 0.75}{\sqrt{\frac{(0.75)(0.25)}{60}}} = -3.036$$

Since this value is outside the two-tailed 5% limits, ± 1.96 , we reject the hypothesis and conclude that the expectation was not correctly estimated. ✓

————— X —————

Problem 12: Two types of computer systems are being tested for use in a new gun. The first system gave 250 hits out of 300 rounds, and the second one gave 182 hits out of 240 rounds. At 1% level can you say that the two systems differ?

Solution: $H_0 : \pi_1 = \pi_2, H_1 : \pi_1 \neq \pi_2$

$$\text{Here } n_1 = 300, n_2 = 240, p_1 = \frac{250}{300}, 0.833, p_2 = \frac{182}{240} = 0.758$$

$$\text{Now, } p = \frac{250 + 182}{300 + 240} = 0.8$$

$$\text{The test statistic } z = \frac{0.833 - 0.758}{\sqrt{(0.8 \times 0.2)\left(\frac{1}{300} + \frac{1}{240}\right)}} = 2.17$$

The two-tailed limits at 1% level are ± 2.58 . Since z lies within these limits, we cannot reject H_0 . Therefore, we conclude that the two systems do not differ.

————— X —————

In the two examples above, we have shown how to test the hypotheses about proportions in two-tailed situations. We are sure you will be able to modify the procedure in case you have to deal with a one-tailed situation. You can check it out by doing these exercises now. We have included two-tailed as well as one-tailed situations here. In case you have a problem, you know you can always find the solution at the end of the unit.

each of size 500 are selected from the bottles produced by the machined. The sample from the first machine was found to contain 250 defective bottles, while that from the second machine contained 40 defective bottles. Is it reasonable to say that both the machines produce the same proportion of defectives if we use $\alpha = 0.05$?

- E17) A random sample of size 1000 from machine 1 contained 20 defectives, and a random sample of size 1500 from machine 2 contained 40 defectives. If $\alpha = 0.05$, can you say that machine 1 is better than machine 2?
- E18) A flue vaccine was given to 125 of a total of 200 employees of a firm. Thirty employees who had received the vaccine were down with flue, while 25 of those who did not, also were stricken. At 1% level of significance would you say that the vaccine was effective?
-

That brings us to the end of this unit. We now summarise our discussion.

6.6 SUMMARY

In this unit we have discussed

- some test procedures that are called significance test or hypothesis test for making inference about the population parameter using sample results. These tests are applied when we have a claim/hypothesis about the population parameter.
 - null and alternate hypothesis, significance level, test statistic, rejection region as the basic elements of these tests.
 - two types of errors - Type 1 and Type 2 errors
 - when to use one-tailed or two-tailed tests.
 - how to apply z-test and t-test.
-

6.7 SOLUTIONS/ANSWERS

- E1) H_0 : It is not a cough medicine.

H_1 : It is a cough medicine.

Suppose it is actually true that it is not a cough medicine i.e. H_0 is true, then if you are rejecting this hypothesis, i.e. you are making a decision that it is a cough medicine. Then this will lead to Type 1 error.

Suppose it is actually true that it is a cough medicine (i.e. H_0 is false) and you are accepting the hypothesis and decide not cough medicine then this will lead to Type II error.

- E2) No.

- E3) The engineer would be interested in whether a bridge of this age could withstand minimum load-bearing capacities necessary for safety purposes. She therefore wants its capacity to be above a certain minimum level, so a one-tailed test would be appropriate. The hypotheses are

$H_0 : \mu = 10$ tons, $H_1 : \mu > 10$ tons

- E4) $H_0 : \mu = 1800$ N

$H_1 : \mu > 1800$ N

$$|z| = \left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right| = \left| \frac{1850 - 1800}{100/\sqrt{50}} \right| = 3.5355 \text{ The critical value for one-tailed test at}$$

5% level is 1.64 and that at 1% level is 2.33. Since $|z|$ is greater than both these, we reject the hypothesis (H_0) and accept the alternative hypothesis (H_1) at both levels. So the new technique is effective.

- E5) $\bar{x} = 0.83$ sec., $s = 0.31$ sec., $n = 300$

The 95% C.I. for μ is

$$\begin{aligned}\bar{x} &\pm 1.96 \frac{s}{\sqrt{n}} \\ &= 0.83 \pm 1.96 \frac{0.31}{\sqrt{300}} = 0.83 \pm 0.035 \\ &= (0.795, 0.865)\end{aligned}$$

With 95% confidence we can say that the mean reaction time of drivers is between 0.795 and 0.865 seconds.

- E6) d.f. = 9 $H_0 : \mu = 23.2\%$

$H_1 : \mu \neq 23.2\%$.

The critical value of t with 9 d.f. at $\alpha = 0.05$ is 2.26

$$|t| = \left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right| = 3.9528$$

Therefore we reject H_0 . The product does not meet the specifications.

- E7) ~~σ~~ = 0.0002 cm., $n = 10$, $\bar{x} = 0.5046$, d.f. = 9 and $\alpha = 0.01$

$H_0 : \mu = 0.51$ cm.

$H_1 : \mu \neq 0.51$ cm.

This is a 2-tailed test.

$$|t| = \left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right| = \left| \frac{0.5046 - 0.51}{0.0002/\sqrt{10}} \right| = 85.3815$$

The critical value of t with 9 d.f. at $\alpha = 0.01$ is 2.821. Hence we reject H_0 .

\therefore the true mean diameter is not 0.51 cm.

- E8) $n_1 = 22$ $n_2 = 18$, $\bar{x}_1 = 127$, $\bar{x}_2 = 136$ $\sigma = 25$ and $\alpha = 0.1$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_2 > \mu_1$

$$|z| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| = \left| \frac{127 - 136}{25 \sqrt{\frac{1}{22} + \frac{1}{18}}} \right| = 1.1327.$$

The critical value of z for $\alpha = 0.1$ is 1.64.

\therefore we do not reject H_0 .

\therefore the new technique is not better.

- E9) $n_1 = 40$ $n_2 = 36$ $\bar{x}_1 = 131$, $\bar{x}_2 = 126$, $\sigma = 16$ and $\alpha = 0.01$

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

$$|z| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| = \left| \frac{131 - 126}{16 \sqrt{\frac{1}{40} + \frac{1}{36}}} \right| = 1.36.$$

The critical value for $\alpha = 0.01$ for a 1-tail test is 2.33.

\therefore We accept H_0

So male students are as smart as female students.

E10) $n_1 = 100$, $n_2 = 75$, $\bar{x}_1 = 1190\text{h}$, $\bar{x}_2 = 1230\text{h}$, $s_1 = 90\text{h}$ $s_2 = 120\text{h}$
and $\alpha = 0.05$

- a) $H_0 : \mu_1 = \mu_2$
 $H_1 : \mu_1 \neq \mu_2$

$$|z| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| = \left| \frac{1190 - 1230}{\sqrt{\frac{8100}{100} + \frac{14400}{75}}} \right| = 2.421.$$

The critical value is 1.96 for a 2-tail test

\therefore we reject H_0
 \therefore There is a significant difference.

- b) $H_0 : \mu_1 = \mu_2$
 $H_1 : \mu_2 > \mu_1$
 $|z| = 2.421$. The critical value is 1.64 for a 1-tail test at $\alpha = 0.05$.
 \therefore we reject H_0 and accept H_1 .
 \therefore The bulbs of manufacturer B are superior to those of manufacturer A.

E11) $n_1 = 12$, $n_2 = 12$, $\bar{x}_1 = 48\text{kg}$, $\bar{x}_2 = 51\text{kg}$, $s_1 = 4\text{kg}$, $s_2 = 3.6\text{kg}$, $\alpha = 0.05$ and d.f. = 22

- $H_0 : \mu_1 = \mu_2$
 $H_1 : \mu_1 > \mu_2$

Now,

$$|t| = \left| \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p = 3.8052 \quad |t| = 1.9312$$

The critical value $t_{0.05} = 1.72$

\therefore we reject H_0 and accept H_1 .
 \therefore There is improvement.

The critical value $t_{0.01} = 2.51$

\therefore At $\alpha = 0.01$, we fail to reject H_0 , and conclude that there is no improvement.

E12)	Top	3	6	7	5	8	9	10	10	7	8	6	9	10	12	4
	Bottom	7	9	5	8	8	10	11	12	6	9	7	13	8	13	8
	Difference	-4	-3	2	-3	0	-1	-1	-2	1	-1	-1	-4	2	-1	-4

$$n = 15, \quad \bar{d} = -1.3333 \quad s = 1.955 \quad \text{d.f.} = 14, \quad \alpha = 0.01$$

- $H_0 : \bar{D} = 0$
 $H_1 : \bar{D} \neq 0$

$$|t| = \left| \frac{\bar{d}}{s/\sqrt{n}} \right| = \frac{1.3333}{1.955/\sqrt{15}} = 2.6412$$

This is a 2-tailed test. $t_{0.01}$ for 14 d.f. is 2.98 \therefore we accept H_0 .

\therefore there is no significant difference.

E13)	Before	68	71	84	93	67	74	82	77	71	83	62	66
	After	71	70	81	97	73	80	90	76	80	79	80	67
	Difference	3	1	3	4	-6	6	-8	1	9	4	18	1

Statistical Inference

$$\bar{d} = 3.833 \quad s_d = 5.9 \quad n = 12 \quad d.f. = 11$$

$$H_0 : \bar{d} = 0$$

$$H_A : \bar{d} < 0$$

$$|t| = \left| \frac{\bar{d}}{s/\sqrt{n}} \right| = \frac{3.833}{5.9/\sqrt{12}} = 2.2505$$

This is a 2-tailed test. The critical value of t for $\alpha = 0.1$ for 11 d.f. is 1.36.

Hence we reject H_0 and accept H_1 .

therefore there is a significant increase in the pulse rate.

$$E14) \sigma = 3 \text{ km/hr.} \quad n = 16, \quad s = 14 \text{ km/hr.} \quad \alpha = 0.05$$

$$H_0 : \sigma = 3 \text{ km/hr.}$$

$$H_A : \sigma > 3 \text{ km/hr.}$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{15(14)^2}{9} = 326.67.$$

This is a 1-tailed test.

$$\chi^2_{0.05} \text{ for } 15 \text{ d.f. is } 25$$

\therefore we reject H_0 and accept H_1 .

$$E15) \sigma^2 = 0.0022 \text{ cm.} \quad s^2 = 0.0031 \text{ cm.} \quad n = 13 \quad \alpha = 0.01.$$

$$H_0 : \sigma^2 = 0.0022 \text{ cm.}$$

$$H_A : \sigma^2 > 0.0022 \text{ cm.}$$

$$\chi^2 = \frac{12(0.0031)}{0.0022} = 16.91$$

$$\chi^2_{0.01} \text{ for } 12 \text{ d.f. is } 26.2.$$

\therefore we accept H_0 . The printer should not reject.

$$E16) n_1 = n_2 = 500 \quad p_1 = \frac{250}{500} = 0.5, \quad p_2 = \frac{40}{500} = 0.08.$$

$$p = \frac{250 + 40}{1000} = 0.29$$

$$H_0 : \pi_1 = \pi_2$$

$$H_A : \pi_1 \neq \pi_2$$

$$|z| = \left| \frac{0.5 - 0.08}{\sqrt{0.29 \times 0.71 \left(\frac{1}{500} + \frac{1}{500} \right)}} \right| = 14.635$$

The critical value for $\alpha = 0.05$ is 1.96. Hence we reject H_0 .

$$E17) n_1 = 1000, \quad n_2 = 1500, \quad \alpha = 0.05$$

$$p_1 = \frac{20}{1000} = 0.02 \quad p_2 = \frac{40}{1500} = 0.0267$$

$$p = \frac{60}{2500} = 0.024$$

$$H_0 : \pi_1 = \pi_2$$

$$H_A : \pi_1 < \pi_2$$

$$|z| = \left| \frac{0.02 - 0.0267}{\sqrt{0.024 \times 0.976 \left(\frac{1}{1000} + \frac{1}{1500} \right)}} \right| = 1.0723$$

1-tailed limit for $\alpha = 0.05$ is 1.64.

\therefore we accept H_0 .

$$E18) n_1 = 125, \quad n_2 = 75, \quad \alpha = 0.01$$

$$p_1 = \frac{30}{125} = 0.24 \quad p_2 = \frac{25}{75} = 0.333$$

$$p = \frac{55}{200} = 0.275$$

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 < \pi_2$$

$$|z| = \left| \frac{0.24 - 0.333}{\sqrt{0.275 \times 0.725 \left(\frac{1}{125} + \frac{1}{75} \right)}} \right| = 1.426$$

This is a 1-tailed test. The critical value of z for $\alpha = 0.01$ is 2.33.

$$|z| < 2.33.$$

\therefore we accept H_0 .

UNIT 7 APPLICATIONS OF CHI-SQUARE IN PROBLEMS WITH CATEGORICAL DATA

Structure		Page Nos.
7.1	Introduction	75
	Objectives	75
7.2	Goodness-of-fit	76
7.3	Test of Independence	82
7.4	Summary	86
7.5	Solutions to Exercises	87
	Appendix : Multinomial Distribution	89

7.1 INTRODUCTION

In this block, you have already studied several problems of testing of hypotheses. The tests that you have studied so far relate to problems where the sample data have been obtained from a continuous distribution, for example, the normal distribution. In practice, however, one often obtains data in which the sampled “observations” are classified into classes according to one or more attributes. For example, a sample of flowers can be classified according to their colour — some of the flowers in the sample could be white, the others could be purple. Again, suppose it is claimed that a vaccine controls a disease. To ‘verify’ the truth of this claim, a sample of N individuals is taken and these individuals can be classified according to two attributes — inoculated or not inoculated, and affected or not affected by the disease.

When the sampled data are classified according to one or more attributes, we say that we have a set of **categorical data**. How do we tackle the inference problems arising out of categorical data? In this unit we shall discuss the use of one of the most widely used tests, the chi-square test, in this context.

To start with, in Sec.7.2, we shall consider the use of the chi-square test in “goodness-of-fit” problems. Then, in Sec.7.3, we shall see how the chi-square test can help us compare two features of a population to see if there is any relationship between them or not. In other words, we test to see if the features occur independent of each other or not.

While studying this unit, please keep comparing the situations in this unit and Units 5 and 6 to really understand the difference in the questions being asked and answered.

Objectives

After studying this unit, you should be able to

- define categorical data;
- identify inference problems associated with categorical data;
- use the chi-square test for solving some inference problems arising in categorical data.

7.2 GOODNESS-OF-FIT

Let us begin by trying to solve Ms.Dalta's problem. She is the marketing director of a company that sells four types of steel almirahs. As part of her duties, she has to make sure that there is no loss of sales due to less stock availability. So far she has been ordering new cupboards assuming that the demand for all four types is the same.

Recently, however, the stock inventories have become more difficult to control. Therefore, Ms.Dalta feels that she should check whether her hypothesis of uniform demand is valid or not.

Can you apply any of the methods you have studied so far for helping Ms.Dalta? There is no parameter that she is estimating and no assumption regarding the distribution of the population. So Ms.Dalta needs to look for some new tools.

What she needs to do is to test the hypothesis :

H_0 : The demand is uniform for all four types of almirahs
against

H_A : The demand is not uniform for all four types of almirahs.

For doing this, she selects a sample of 80 almirahs sold over the past few months. Ms.Dalta assumes that the demand is uniform. So the probability of an almirah of Type i being bought is the same, for $i = 1, 2, 3, 4$. If we denote this probability by p_i , then $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$. So, if the demand is uniform, she can expect

$80 \left(\frac{1}{4} \right) = 20$ almirahs of each type to be sold. But the observed sales of each

type are 23, 19, 18 and 20, respectively. Her problem is to see how well her hypothesis of uniform demand fits the observed sales. In other words, how can this data set be used for testing H_0 ?

More generally, suppose a sample of n individuals are classified into k classes. Suppose the number of individuals falling in the i th class is O_i ($i = 1, \dots, k$).

The problem of “**goodness-of-fit**” consists in testing the hypothesis, H_0 , that the probability of an individual falling in the i th class is p_i ($i = 1, \dots, k$), where

$\sum_{i=1}^k p_i = 1$. In other words, the hypothesis H_0 to be tested is that the number of individuals (in a sample of size n) falling in the i th class is np_i ($i = 1, \dots, k$). This is to be tested against the hypothesis H_A , that H_0 is not true.

So, in this general situation, the “expected” number of individuals falling in the i th class is $E_i = np_i$ ($i = 1, \dots, k$). Note that these “expected” numbers are known to us because to start with we assume H_0 and calculate them. That is, we assume that the probability of an individual falling in the i th class is p_i ($i = 1, \dots, k$). Based on these numbers $O_1, O_2, \dots, O_k, E_1, E_2, \dots, E_k$, there is a way of testing the validity of H_0 . Let us see what this method is.

Let $U = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$. This statistic U , under some mild conditions, is known to

have an **approximate chi-square distribution with $k - 1$ degrees of freedom**, where k is the number of classes. If we want to test the hypothesis H_0 at the α level of significance, then we need to find $\chi^2_{\alpha, k-1}$, from the standard χ^2

distribution tables (given at the end of this block). If $U > \chi^2_{\alpha, k-1}$, we reject H_0 .

Otherwise we do not reject H_0 .

H_0 is the null hypothesis, and H_A is the alternative hypothesis.

Note that, $\chi^2_{\alpha, k-1}$ is denoted by ‘ χ^2_α with $k-1$ degrees of freedom’ in Unit 4.

U is also called the sample χ^2 value, or the observed value of χ^2 for the data.

To see how this test works, let us consider Ms.Dalta's data, presented in Table 1.

Table 1

Type of almirah	Observed sales (O_i)	Expected sales (E_i) $= np_i = 80 \times \left(\frac{1}{4}\right)$
I	23	20
II	19	20
III	18	20
IV	20	20

$$\text{Here } U = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

$$= \frac{(23 - 20)^2}{20} + \frac{(19 - 20)^2}{20} + \frac{(18 - 20)^2}{20} + \frac{(20 - 20)^2}{20}$$

$$= \frac{9 + 1 + 4 + 0}{20} = 0.7$$

Here $k = 4$. If Ms.Dalta wants to test H_0 at a 5% level of significance, $\alpha = 0.05$. Now, $\chi^2_{0.05,3} = 7.815$. Since $U < \chi^2_{0.05,3}$, Ms.Dalta does not reject H_0 . In other words, Ms.Dalta concludes that the demand for the four types of almirahs is uniform.

Another example may help you to see how this test works.

Example 1 (Experiment on the breeding of flowers of a certain species) :

Jaswant is interested in breeding flowers of a certain species. The experimental breeding can result in four possible types of flowers :

- (a) magenta flowers with a green stigma (MG),
- (b) magenta flowers with a red stigma (MR),
- (c) red flowers with a green stigma (RG),
- (d) red flowers with a red stigma (RR).

According to the well-known Mendel's law, these four kinds of flowers should come out in the ratio 9 : 3 : 3 : 1. Jaswant found that under her experiment, out of 160 flowers that bloomed, the number of flowers with types MG, MR, RG and RR were 84, 35, 28 and 13, respectively. She wants to find out whether these data are compatible with Mendel's law or not.

If they are compatible, then the probabilities of each of these types of blooming are $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, and $p_4 = \frac{1}{16}$. So Jaswant wants to test the hypothesis

H_0 : The distribution of the flower types is multinomial with

$$p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}.$$

against

H_A : H_0 is not true, that is, the distribution is not multinomial with the specified probabilities.

Jaswant's data can be presented as shown in Table 2.

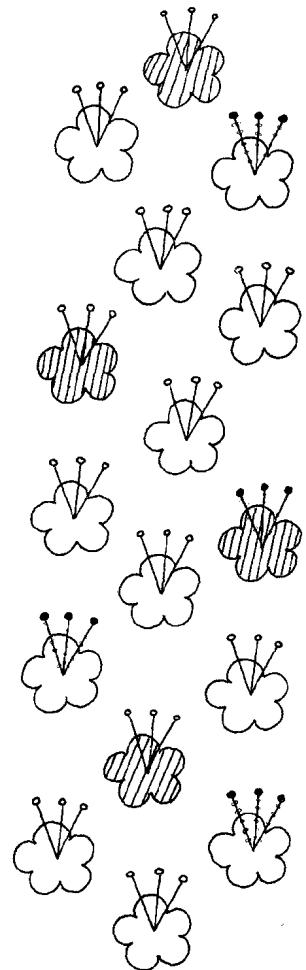


Fig.1

See the appendix to the unit for a brief introduction to the multinomial distribution.

Table 2

Flower type	Observed number O_i	Expected number $E_i (= np_i)$
MG	84	90
MR	35	30
RG	28	30
RR	13	10
Total (n)	160	160

Here $k = 4$, and

$$\begin{aligned} U &= \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(84-90)^2}{90} + \frac{(35-30)^2}{30} + \frac{(28-30)^2}{30} + \frac{(13-10)^2}{10} \\ &= 2.27 \end{aligned}$$

Jaswant needs to compare this value with the appropriate critical χ^2 -value. She takes the significance level of the test as $\alpha = 0.05$. Also, in this case, since the number of classes is 4, the degrees of freedom are $4 - 1 = 3$. So, she finds

$\chi^2_{0.05,3}$, which is 7.81. Since $U = 2.27 < 7.81 = \chi^2_{0.05,3}$, she does not reject H_0 .

Thus, Jaswant concludes that her data is compatible with Mendel's law.

* * *

In the two situations above, the hypothetical probabilities p_1, p_2, \dots were known to us from before because of the type of assumption H_0 was. However, in some problems, these probabilities may have to be estimated from the data itself. The following example illustrates this.

Example 2 : A consultant was employed by a city council to study the pattern of bus arrivals and departures at a very busy interstate bus terminus. Since many

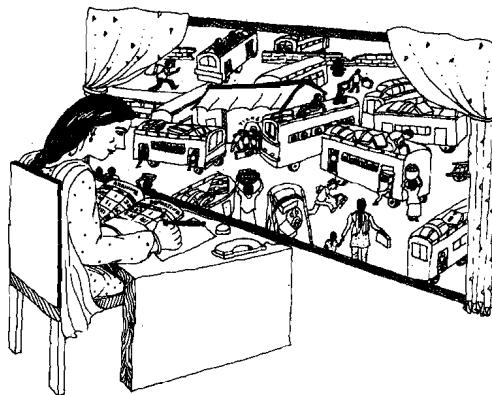


Fig.2

arrival processes fit the Poisson distribution, she decided to test the following hypothesis :

H_0 : The arrivals are distributed as a Poisson random variable,
against

H_A : The arrivals are not Poisson distributed.

She sampled the number of arrivals in 200 minutes. Then she grouped the arrivals into $k = 6$ categories, and noted her observations, as shown in Column 2 of Table 3 below.

Applications of Chi-Square in Problems with Categorical Data

However, since the parameter of the Poisson distribution is unspecified in the hypothesis, the consultant needed to estimate this from the data itself. For this she first computed the sample mean as

$$\bar{x} = \frac{(1 \times 23) + (2 \times 45) + \dots + (5 \times 41)}{200}$$

$$= 2.96$$

So, she estimated the parameter of the Poisson distribution as $\hat{\lambda} = 2.96$.

With this value of $\hat{\lambda}$, she computed the Poisson probabilities for the different classes from the tables (which are also provided at the end of this block). These are shown in Column 3 of the table below.

Table 3 : Arrivals at ISBT

No. of arrivals	Observed frequencies O_i	Prob. according to Poisson dist. p_i	Expected frequencies $E_i (=np_i)$
0	10	0.0524	10.48
1	23	0.1545	30.90
2	45	0.2277	45.54
3	49	0.2238	44.76
4	32	0.1651	33.02
5 or more	41	0.1765	35.30
Total	200	1.0000	200.00

According to her data,

$$U = \frac{(10 - 10.48)^2}{10.48} + \frac{(23 - 30.90)^2}{30.9} + \dots + \frac{(41 - 35.3)^2}{35.3}$$

$$= 0.022 + 2.02 + 0.006 + 0.402 + 0.032 + 0.92$$

$$= 3.402$$

Here $k = 6$ but one parameter has been estimated. So, the degrees of freedom associated with the chi-square distribution is $(k - 1) - 1 = k - 2 = 4$. The critical value of chi-square at 4 degrees of freedom and 1 percent level of significance is 13.27. Since $3.402 < 13.27$, the consultant did not reject the null hypothesis. In other words, she was in a position to conclude that the arrivals and departures at the bus terminus were Poisson distributed.

* * *

Let us now look at a problem involving normal distribution. While solving it, the following very important point about applying the χ^2 -test will show up.

To find p_i for $\lambda = 2.96$, we take the average of the values in the columns corresponding to 2.9 and 3.0, respectively. Thus,

$$p_1 = \frac{0.055 + 0.0498}{2} = 0.0524.$$

Remark 1 : If, corresponding to a category, say j , the expected value E_j is small, i.e., less than 5, then the chi-square approximation for the distribution of U will not be good. So, if the condition $E_i \geq 5$ is not satisfied for all i , then we should combine the category j with $E_j < 5$ with its adjacent categories $j + 1, j + 2, \dots, j + r$, where $E_j + E_{j+1} + \dots + E_{j+r} \geq 5$ but $E_j + E_{j+1} + \dots + E_{j+r-1} < 5$. The number of classes, accordingly, gets reduced by r .

This remark will become more clear as you study the solution of Problem 1.

Problem 1 : A chemical company wants to know if its sales of a liquid chemical are normally distributed. This information will help them in planning and

controlling the inventory. The sales record for a random sample of 200 days is given in Table 4.

Table 4

Sales (in 1000 litres)	Number of days
Less than 34.0	0
34.0-35.5	13
35.5-37.0	20
37.0-38.5	35
38.5-40.0	43
40.0-41.5	51
41.5-43.0	27
43.0-44.5	10
44.5-46.0	1
46.0 or more	0
Total	200

We assume that the upper limit of a class shows that quantities less than that limit are in the class. So, for example, 35.5 will be included in the third class interval, not the second one.

At the 5% level of significance, test the hypothesis that the company's sales are normally distributed.

Solution : Let us start by clearly stating our hypotheses.

H_0 : The company's sales are normally distributed.
against

H_A : The company's sales are **not** normally distributed.

Now, we assume for just now that H_0 is valid. By methods known to us, we can calculate the sample mean and sample standard deviation \bar{x} and s_x . You can check that these are :

$$\bar{x} = 40,000 \text{ litres}, s_x = 2.5 \text{ thousand litres.}$$

Now, we need to find the expected frequencies E_i corresponding to each O_i . You know that $E_i = 200 \times p_i$, where p_i is the probability for each class in Table 4, computed under the assumption of normal distribution.

So, let us expand Table 4 to include all the class probabilities (Column 3), the expected frequencies (Column 4) and the corresponding values of $\frac{(O_i - E_i)^2}{E_i}$ (Column 5).

To get the first entry in Column 3, we compute $z = \frac{(x - \mu)}{\sigma}$ for $x = 34$. As you know, μ and σ are estimated by \bar{x} and s_x , respectively. So, $z = \frac{(34 - 40)}{2.5} = -2.4$.

Now, from the table of normal probabilities in the Block Appendix, you know that $P[-2.4 \leq Z \leq 0] = P[0 \leq Z \leq 2.4] = 0.4918$.

So, the probability we want is

$$p_1 = 0.5 - P[-2.4 \leq Z \leq 0] = 0.5 - 0.4918 = 0.0082.$$

$$\text{Therefore, } E_1 = 200(0.0082) = 1.64.$$

Similarly, you can compute the other expected frequencies and complete the 4th column of Table 5. You may wonder about the brackets in Columns 2, 3 and 4 of the table. This is because, as we have mentioned in Remark 1, **the χ^2 goodness-of-fit test is a good approximation only if the E_i are not very small**. This is why we have grouped the first two classes and the last two classes in Table 5.

To fill in the fifth column of Table 5, we treat the bracketed classes as a single

class. So, $\frac{(O_1 - E_1)^2}{E_1} = \frac{(13 - 7.18)^2}{7.18} = 4.7176$. You can similarly calculate the

other entries of Column 5 in the table below.

Table 5

**Applications of Chi-Square
in Problems with Categorical
Data**

Sales (in 1000 litres)	Observed frequency (O_i)	Class probability (p_i)	Expected frequency (E_i)	$\frac{(O_i - E_i)^2}{E_i}$
less than 34.0	0	0.0082	1.64	
34.0 – 35.5	13	0.0277	5.54	7.18
35.5 – 37.0	20	0.0792	15.84	1.0925
37.0 – 38.5	35	0.1592	31.84	0.3136
38.5 – 40.0	43	0.2257	45.14	0.1015
40.0 – 41.5	51	0.2257	45.14	0.7607
41.5 – 43.0	27	0.1592	31.84	0.7357
43.0 – 44.5	10	0.0792	15.84	2.1531
44.5 – 46.0	1	0.0277	5.54	
greater than 46.0	0	0.0082	1.64	7.18

Now, summing up the entries in the last column of Table 5, we get $U = 15.194$.

Next, to see whether we accept or reject H_0 , we look up the value of χ^2 at the 5% level of significance and for the appropriate number of degrees of freedom. Note that, though we started with the data categorised into 10 classes, we needed to group two sets of 2 frequencies each. So, for purposes of the χ^2 test we now have 8 classes. Also, we have estimated two parameters, μ and σ . Therefore, the degrees of freedom are $(8 - 1) - 2 = 5$.

So, from the χ^2 table, we find $\chi^2_{0.05,5} = 11.07$.

Since $U > \chi^2_{0.05,5}$, we must reject H_0 . That is, the normal distribution is not a good fit to the data.

* * *

Now try the following exercises.

- E1) In Table 6 below you find the distribution of the heights for 100 college students. Estimate the mean and the standard deviation of the distribution. Check whether the sample is drawn from a normally distributed population at 5% level of significance.

Table 6

Class (cm)	Number of students (O_i)
Less than 161	4
161 - 164	11
164 - 167	16
167 - 170	19
170 - 173	25
173 - 176	18
176 - 179	4
179 - 182	2
182 or more	1
Total	100

- E2) Test whether the observed frequencies, as given below, in 4 phenotypic classes AB, Ab, aB, ab are in agreement with the expected ratio 9: 3: 3: 1.

Class	AB	Ab	aB	ab
Frequency	102	25	28	5

E3) A die is rolled 1200 times with the following results :

No. that comes up	1	2	3	4	5	6
Frequency	205	279	217	257	133	109

Test if the die is unbiased.

In all the situations so far, the problem was related to data that were classified according to one attribute. Now let us see how the χ^2 test can be used to infer about situations in which the data are classified according to two or more attributes.

7.3 TEST OF INDEPENDENCE

In this section we shall look at inference problems like the following one.

Dr.Surya had recently developed a serum that she thought might be effective in preventing colds. But, she needed to verify its efficacy. For this purpose she carried out an experiment.

One thousand individuals were classified into two groups of the same size. The serum was administered to the members of the first group only. The number of individuals in each group who caught a cold zero times, or once, or more than once during some period after the treatment was noted. The data are shown in the following table having 2 rows and 3 columns.



Fig.3 :“ Don’t worry!
You take this
medicine, and you
won’t have any more
colds in future.”

Table 7 : Table showing the effect of serum

Category	The number catching a cold			Total
	Zero times	Once	More than once	
Group given serum	252	145	103	500
Untreated group	224	136	140	500
Total	476	281	243	1000

Dr.Surya’s problem was to examine whether or not this serum is effective in preventing a cold. In other words, she wants to know if a person can catch a cold one or more times whether s/he has taken the serum or not. We can reword this as: is the treatment by the serum **independent** of the number of times of catching a cold?

So, Dr.Surya formulated the following null and alternative hypotheses to be tested:

H_0 : There is no interdependence between the serum treatment and the number of times of getting a cold.

H_A : H_0 is not true, i.e., the serum has some effect on preventing colds.

To test H_0 against H_A she planned to use the χ^2 test at the 5% significance level. As you know, to do so she needed to calculate the expected frequencies corresponding to each of the 6 entries in the 2×3 table, Table 7, assuming H_0 , i.e., the independence of the number of times one gets a cold and of taking serum treatment.

Let us see how she obtained E_{11} . For this, she used the fact that out of the 1000 people, 476 had no cold. So, out of the 500 in the treatment group,

$\frac{476}{1000} \times 500 = 238$ were expected to not have any cold. Note that this is

$$\frac{(\text{sum of the first row entries}) \times (\text{sum of first column entries})}{(\text{total sample size})}$$

Similarly, she calculated the other expected frequencies :

$$E_{12} = 500 \times \frac{281}{1000} = 140.5, E_{13} = 121.5, E_{21} = 238, E_{22} = 140.5, E_{23} = 121.5.$$

Then , Surya calculated the sample statistic U as

$$U = \frac{(252 - 238)^2}{238} + \frac{(145 - 140.5)^2}{140.5} + \frac{(103 - 121.5)^2}{121.5} + \frac{(224 - 238)^2}{238} \\ + \frac{(136 - 140.5)^2}{140.5} + \frac{(140 - 121.5)^2}{121.5} = 7.57$$

She took the significance level of the test as $\alpha = 0.05$. Also, in this case the number of degrees of freedom was $(2 - 1)(3 - 1) = 2$. So, comparing the value of U with $\chi^2_{0.05, 2} = 5.99$, she found that $U > \chi^2_{0.05, 2}$

So, she rejected H_0 , and concluded that the serum has some effect in preventing colds.

For an $m \times n$ table, the number of degrees of freedom is $(m-1)(n-1)$.

Let us look closely at the steps Dr. Surya went through for testing the independence of two features of the population under study.

Step 1 : She stated the hypothesis regarding the independence of two features of the sample.

Step 2 : As in the case of the goodness-of-fit tests, she noted the frequencies — how many of each type of person (treated or untreated) had which kind of feature (the number of times they catch a cold). These frequencies were written in a table, called a **contingency table**.

In this case, the contingency table had 2 rows and 3 columns, because corresponding to each of the two groups of people there were 3 possibilities about the cold they did or did not catch. In brief, we say that the table was a 2×3 contingency table.

Step 3 : Corresponding to each of the 6 cells of the contingency table, Dr. Surya calculated the expected frequency. She did this as follows :

$$E_{ij} = \text{expected frequency for } i\text{th row and } j\text{th column} \\ = \frac{(\text{sum of entries of } i\text{th row})(\text{sum of entries of } j\text{th column})}{(\text{total sample size})}$$

where $i = 1, 2$ and $j = 1, 2, 3$.

Step 4 : Then the sample χ^2 , U, was calculated by

$$U = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ where } O_{ij} \text{ was the entry in the } i\text{th row and } j\text{th column.}$$

Note that, more generally, if she had had an $m \times n$ contingency table, the value would be

$$U = \sum_{i=1}^m \sum_{j=1}^n \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ where } m \text{ and } n \text{ are natural numbers.}$$

Step 5 : She compared this value with the value of $\chi^2_{\alpha,d}$, where α is the level of significance and $d = (m - 1)(n - 1)$ is the number of degrees of freedom. Then, as you saw in Sec. 7.2, if $U < \chi^2_{\alpha,d}$, H_0 is accepted. Otherwise H_0 is rejected.

Another example may help you to clarify your understanding regarding the process of testing for independence.

Example 4 : The Glorious Watch Company wants to find out if there is any relationship between the income of a person and the importance she attaches to the price of a brand name. Mr.Zafar, the Chief of the Marketing Division, wants to test the hypothesis

H_0 : Income of a person and importance to her of price attached are independent.
against
 H_A : H_0 is not true.

Zafar does a survey among the customers. To analyse his results, he groups them into 3 income levels, and asks them to mark the level of importance they give on a 3-point scale — great, moderate or low. He noted the results in a contingency table (see Table 8). In this table, you will also find the expected frequency corresponding to each observed frequency written alongside. As you know, these will be calculated as follows :

$$\begin{aligned} E_{11} &= \frac{(\text{sum of entries of first row})(\text{sum of entries of first column})}{(\text{total sample size})} \\ &= \frac{170 \times 187}{500} = 63.58. \end{aligned}$$

All the other E_{ij} s are calculated in the same way.

Table 8

Feature 1 (Importance Level)	Feature 2 (Income)						Total
	Low		Middle		High		
	O_{i1}	E_{i1}	O_{i2}	E_{i2}	O_{i3}	E_{i3}	
Great	79	63.58	58	61.2	33	45.22	170
Moderate	48	59.09	65	56.88	45	42.03	158
Low	60	64.33	57	61.92	55	45.75	172
Total	187		180		133		500

So, the sample χ^2 value that Zafar calculated was

$$\begin{aligned} U &= \frac{(79 - 63.58)^2}{63.58} + \frac{(58 - 61.2)^2}{61.2} + \dots + \frac{(57 - 61.92)^2}{61.92} + \frac{(55 - 45.75)^2}{45.75} \\ &= 3.74 + 0.167 + 3.302 + 2.081 + 1.159 + 0.21 + 0.291 + 0.391 + 1.87 \\ &= 13.211 \end{aligned}$$

Then Zafar compared this with the value of χ^2 for $(3-1)(3-1) = 4$ degrees of freedom and at the 2% level of significance, which is $\chi^2_{0.02,4} = 11.668$.

Applications of Chi-Square in Problems with Categorical Data

He found $U > \chi^2_{0.02,4}$, which made him decide that he should reject H_0 . In other words, Zafar is 98% certain that the level of income of a person is related to the importance she gives to the price of the brand of watches.

* * *

In the example above, it is interesting to note that if Zafar had chosen to be 99% certain, then $\chi^2_{0.01,4} = 13.277 > U$. So that, he would not have rejected H_0 . What does this tell us about statistical analyses? Think about it.

And now here are some problems for you to solve.

- E4) The data in the following table give mortality rates among vaccinated and non-vaccinated patients. Test if the vaccine has any effect in curing the disease.

Categories	Living	Dead	Total
Vaccinated	320	125	445
Non-vaccinated	98	230	328
Total	418	355	773

- E5) Do the following data on sociability of soldiers recruited in cities and villages suggest that city soldiers are more sociable than village soldiers?

Sociability Place	Sociable	Non-sociable
City	13	6
Village	7	14

- E6) A group of 1650 school children were classified according to their performance in school tests and family economic level. Test if there is any association between these two attributes.

Performance Economic level	Very Good	Good	Average	Poor	Total
Very Rich	4	7	16	25	52
Rich	13	37	79	73	202
Average	105	372	298	175	950
Poor	36	213	75	123	446
Total	157	629	468	396	1650

- E7) In an experiment to study whether smoking affects health, the following data were collected. Test the hypothesis that smoking does not affect health.

	Light smoking	Moderate smoking	Heavy smoking
Health affected	16	29	35
Health not affected	36	23	17

In this section you have seen situations in which the population is tested to see if two or more common features of the population are related or not. This is as far as we intend to discuss the use of χ^2 for analysing categorical data. Let us end with a brief look at what we have covered in this unit.

7.4 SUMMARY

In this unit we have started with a look at data presented in the form of frequencies falling in different categories or classes. Based on such data we have undertaken different tests of hypotheses using the chi-squared distribution. We have considered two types of tests :

- 1) **Test of goodness-of-fit** : The hypotheses are given by
 H_0 : The data fit a given distribution ; against
 H_A : H_0 is not true, i.e., the data do not fit that distribution.

For testing whether H_0 is acceptable, we consider the observed and expected frequencies of the various categories in the data.

Suppose there are k categories with O_i as the observed frequency and E_i as the expected frequency of the i th category. Then the sample χ^2 value is

$$U = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

If H_0 were acceptable, then this value should be less than $\chi^2_{\alpha, k-s-1}$ with 100α % significance level, where s is the number of parameters estimated in finding the expected frequencies.

So, if $U < \chi^2_{\alpha, k-s-1}$, then H_0 is not rejected. Otherwise, H_0 is rejected.

- 2) **Test of independence** : Suppose a population can be classified into r categories on the basis of feature A, and into c categories on the basis of feature B. The hypotheses are given by :

H_0 : There is no interdependence between the features A and B
 H_A : H_0 is not true, that is, A has an effect on B.

The data is presented in the form of an $r \times c$ contingency table. Let n_{ij} be the frequency in the i th row and j th column and let

$$n_{i0} = \sum_j n_{ij}, \quad n_{0j} = \sum_i n_{ij}, \quad n = \sum_i \sum_j n_{ij}$$

If the two classification criteria are mutually independent, the expected value E_{ij} for the i th row and j th column is given by

$$E_{ij} = \frac{n_{i0} \times n_{0j}}{n}$$

Then, the sample χ^2 value, $U = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$.

If this value is less than $\chi^2_{\alpha, (r-1)(c-1)}$, then H_0 is acceptable at the α level of significance.

And now you may like to check whether you have achieved the objectives of the unit listed in Sec.7.1. Also, while doing the exercises in this unit, you may have had some doubts. If so, please go through the following section also.

7.5 SOLUTIONS/ANSWERS

E1) Here $\bar{x} = 170$, $s^2 = 36$ and $n = 100$.

H_0 : The sample is drawn from a population with normal distribution $N(170.0, 6^2)$.

H_A : H_0 is not true.

In order to solve this problem by the same method as in Example 1, we consider the classes in Table 6 corresponding to categories of a multinomial distribution. Let O_i be the observed value for the i th category. Then, what is the expected value for the i th category in this case? Since the population distribution is completely specified as $N(170, 6^2)$ under the null hypothesis H_0 , we can obtain the probability p_i with which the height of a student chosen randomly falls into the i th category. The expected value for the i th category is obtained by $E_i = np_i$. To compute the values p_i , the boundary points of the classes should be standardised by the population means and the standard deviation so as to make use of the table for a standard normal distribution. The standardised boundary points are given below in Table 11.

Table 11

Boundary Points of Class (x_i)	161	164	167	170	173	176	179	182
Standardised Boundary Points (z_i)	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0

Here, $z_i = \frac{x_i - 170.0}{6.0}$ The p_i and E_i values can be obtained as follows.

$$p_1 = P[-\infty < Z < -1.5] = 1 - P[-\infty < Z < 1.5] = 0.0668; E_1 = 100 \times 0.0668 = 6.68$$

$$p_2 = P[-1.5 \leq Z < -1.0] = 0.0919; E_2 = 9.19, \text{ and so on.}$$

The values are all given in Table 12 below.

Table 12

Class (cm)	Number of students (O_i)	Probabilities (p_i)	Expected values E_i ($=np_i$)
Less than 161	4	0.0668	6.68
161 - 164	11	0.0919	9.19
164 - 167	16	0.1498	14.98
167 - 170	19	0.1915	19.15
170 - 173	25	0.1915	19.15
173 - 176	18	0.1498	14.98
176 - 179	4	0.0919	9.19
179 - 182	2	0.0440	4.40
182 or more	1	0.0228	2.28
Total	100	1.0000	6.68

Let us now test H_0 against H_A at the 5% significance level.

Now, from Table 12,

$$\begin{aligned}
 U &= \frac{(4-6.68)^2}{6.68} + \frac{(11-9.19)^2}{9.19} + \frac{(16-14.98)^2}{14.98} + \frac{(19-19.15)^2}{19.15} + \\
 &\quad \frac{(25-19.15)^2}{19.15} + \frac{(18-14.98)^2}{14.98} + \frac{(4-9.19)^2}{9.19} + \frac{(3-6.68)^2}{6.68} \\
 &= 8.86
 \end{aligned}$$

Taking the significance level of the test $\alpha = 0.05$, we have $\chi^2_{0.05,5} = 11.07$.

The degrees of freedom $8 - 3 = 5$, because the number of categories, after combining the last two categories is 8, and the number of parameters estimated is 2.

Since $U = 8.86 < 11.07 = \chi^2_{0.05,5}$, we conclude that there is good agreement between the observed frequencies and the fitted values. So H_0 is accepted.

- E2) If the data are compatible with the given ratios, the expected frequencies are : AB: 90, Ab: 30, aB: 30, ab: 10.

The value of U is 5.07. This is less than $\chi^2_{0.05,3} = 7.815$. Hence, we accept the null hypothesis that the given data are in agreement with the expected ratios.

- E3) Here H_0 : the expected frequency is 200 in each class.
 H_A : H_0 is not true.

Therefore, $U = 112.87 > \chi^2_{0.05,5} = 11.070$. Hence, we conclude on the basis of the given data that we reject H_0 . So the die is not unbiased.

- E4) The hypothesis here is:

H_0 : There is no effect of the vaccine on mortality.
against
 H_A : H_0 is not true.

The expected frequencies E_{ij} are given in the table below.

	Living	Dead	Total
Vaccinated	241	204	445
Non-Vaccinated	177	151	328
Total	418	355	773

The observed value of χ^2 is $U = 133.08$.

The number of degrees of freedom $= (2 - 1)(2 - 1) = 1$.

$$\chi^2_{0.05,1} = 3.84 < U.$$

Hence, we conclude that we cannot accept H_0 . So, on the basis of the given data, we conclude that the vaccine has a definite effect on the mortality rate.

- E5) H_0 : There is no interdependence between place and sociability level.
 H_A : H_0 is not true.

The table of expected frequencies is

Sociability Place	Social	Non-social	Total
City	9.5	9.5	19
Village	10.5	10.5	21
Total	20	20	40

$$\text{So, } U = 12.25 \left(\frac{2}{9.5} + \frac{2}{10.5} \right) = 4.9.$$

The number of degrees of freedom = 1.

$$\chi^2_{0.05, 1} = 3.84 < U.$$

Therefore, we reject H_0 . So, the data suggests that the place a soldier comes from affects her/his sociability level.

- E6) The expected frequencies are given below :

Performance		Very Good	Good	Average	Poor
Economic level	Very Rich	4.95	19.82	14.75	12.48
	Rich	9.22	77.00	57.29	48.48
Average	90.39	362.15	269.45	228	
Poor	42.44	170.02	126.50	107.04	

The value of the sample χ^2 is $U = 127.61 > 25.0 = \chi^2_{0.05, 15}$. Hence, the hypothesis of independence between the categories is rejected.

- E7) Under the assumption that smoking does not affect health, the expected frequencies are given below.

	Light smoking	Moderate smoking	Heavy smoking
Health affected	26.67	26.67	26.67
Health not affected	25.33	25.33	25.33

The observed value of χ^2 is $U = 14.52 > \chi^2_{0.05, 2} = 5.991$. Hence, it is concluded on the basis of the given data that smoking affects health.

APPENDIX : MULTINOMIAL DISTRIBUTION

This distribution is an extension of the binomial distribution that you studied in Unit 3. It shows up in the following situation :

There is an experiment which consists of n identical trials, which are independent. Each trial can have k possible outcomes. Suppose the probability of each of these outcomes is p_1, p_2, \dots, p_k , with $p_1 + p_2 + \dots + p_k = 1$. These probabilities remain the same from trial to trial.

Mathematically, this situation is represented by considering k random variables X_1, \dots, X_k with probabilities p_1, \dots, p_k that $X_1 = x_1, \dots, X_k = x_k$, respectively, where

$\sum_{i=1}^k x_i = n$, $\sum_{i=1}^k p_i = 1$, $p_i \neq 0 \forall i = 1, \dots, k$. If the random vector (X_1, \dots, X_k) is

multinomially distributed, the $P[X_1 = x_1, \dots, X_k = x_k] = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$.

So, if we are testing if a certain population is multinomially distributed, we will test

H_0 : The population of size n is multinomially distributed with probabilities

p_1, p_2, \dots, p_k (known to us);

against

H_A : The population is not multinomially distributed.

As in all the other cases discussed in the unit, if the np_i are not very small, then the

test statistic $U = \sum_{i=1}^k \frac{(O_i - np_i)^2}{np_i}$ has approximately a chi-squared distribution with

$(k - 1)$ degrees of freedom. The approximation is usually good for $E_i = np_i \geq 5$.

APPENDIX – 1

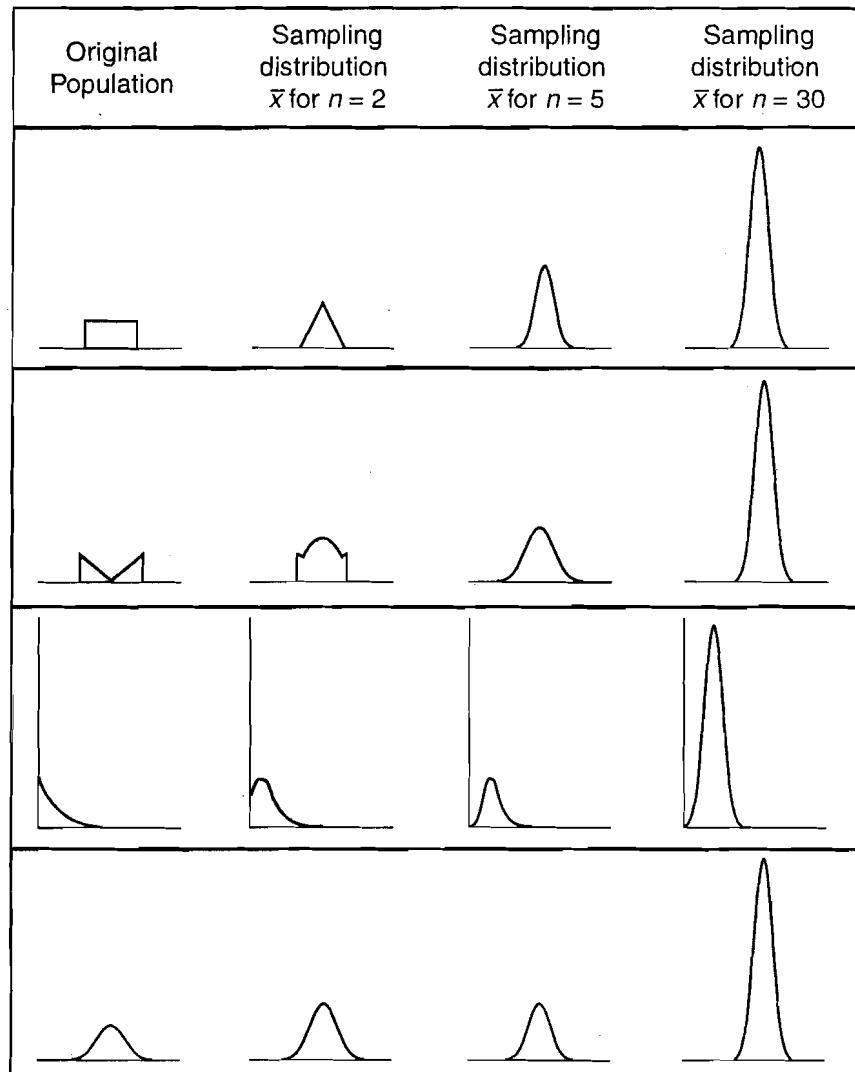
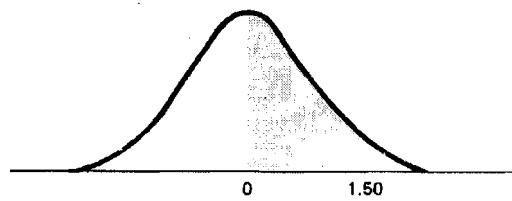


Fig. 1 Sampling distribution of \bar{x} for different populations and different sample sizes.

APPENDIX-2

TABLE 1
AREAS UNDER THE STANDARD NORMAL CURVE

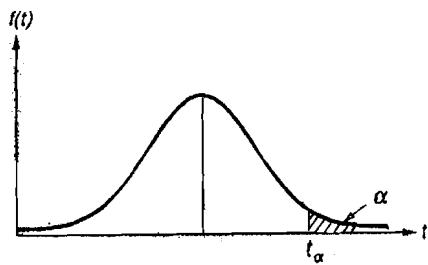
This table shows the area between zero (the mean of a standard normal variable) and z . For example, if $z = 1.50$, this is the shaded area shown below which equals .4332.



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: This table is adapted from National Bureau of Standards, *Tables of Normal Probability Functions*, Applied Mathematics Series 23, U.S. Department of Commerce, 1953.

TABLE-2
t-distribution



ν	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

TABLE-3
CHI-SQUARED DISTRIBUTION

ν	$\alpha = .995$	$\alpha = .990$	$\alpha = .975$	$\alpha = .950$	$\alpha = .900$	$\alpha = .750$	$\alpha = .500$	$\alpha = .250$	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .010$	$\alpha = .005$
1	0.0000393	0.000157	0.000982	0.00393	0.01579	0.10153	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100	13.70069	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032	14.84540	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.33976	15.98391	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.33850	19.36886	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	13.67529	17.33790	21.60489	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.56200	18.33765	22.71781	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	16.34438	20.33723	24.93478	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.13730	22.33688	27.14134	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	21.74940	26.33634	31.52841	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297	46.97924	50.89218	53.67196

TABLE-4
F-distribution

$F_{0.05}$

$v_2 = \text{Degrees of freedom for denominator}$		$v_1 = \text{Degrees of freedom for numerator}$																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.50	19.00	19.20	19.30	19.40	19.40	19.40	19.40	19.40	19.40	19.40	19.40	19.40	19.50	19.50	19.50	19.50	19.50	19.50
3	10.10	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.38	2.38	2.30	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	3.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.93
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

TABLE-5
F-distribution

$F_{0.01}$

$\nu_2 = \text{Degrees of freedom for denominator}$		$\nu_1 = \text{Degrees of freedom for numerator}$																	
ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4.052	5.000	5.403	5.625	5.764	5.859	5.928	5.982	6.023	6.056	6.106	6.157	6.235	6.287	6.313	6.339	6.366		
2	98.50	99.00	99.20	99.30	99.40	99.50	99.50	99.40	99.40	99.40	99.40	99.50	99.50	99.50	99.50	99.50	99.50	99.50	
3	34.10	30.80	29.50	28.70	28.20	27.90	27.50	27.30	27.20	27.10	26.90	26.70	26.50	26.40	26.30	26.20	26.10		
4	21.20	18.00	16.70	16.00	15.50	15.20	15.00	14.80	14.70	14.50	14.40	14.20	14.00	13.90	13.70	13.60	13.50		
5	16.30	13.30	12.10	11.40	11.00	10.70	10.50	10.30	10.20	10.10	9.89	9.72	9.55	9.47	9.38	9.20	9.11	9.02	
6	13.70	10.90	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.88	
7	12.20	9.53	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	
8	11.30	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	
9	10.60	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	
10	10.00	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.60	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.36	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.17	
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.06	2.87	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.75	
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83		
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.15	3.08	3.00	2.92	2.84	2.75	
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.06	2.92	2.84	2.76	2.67	2.49	
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.42	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.36	
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.31	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.55	2.26	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.21	
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.17	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.01	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	1.92	1.80	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.73	1.60	
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.53	1.38	
∞	6.63	4.61	3.76	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	

UNIT 8 ANALYSIS OF VARIANCE: ONE-WAY CLASSIFICATION

Structure	Page No.
8.1 Introduction	5
Objectives	
8.2 Analysis of Variance: Basic Concepts	6
Source of Variance	
One-Way Classification	
8.3 Model For One-Way Classification	8
8.4 Test Procedure	9
Sums of Squares	
Preparation of ANOVA Table	
8.5 Pairwise Comparisons	14
8.6 Unbalanced Data	15
8.7 Random Effects Model	17
8.8 Summary	19
8.9 Solutions/Answers	19

8.1 INTRODUCTION

In Unit 6, you have learnt to compare two population means. Often, in practice, one is required to compare more than two population means. In this unit, we shall study a statistical procedure, called analysis of variance, which allows one to test a hypothesis comparing several normal population means.

The problem of comparing several population means arises quite naturally in practice. For instance, on the basis of sample data, one might wish to decide whether there is any real difference between three teaching methods of a foreign language. Quite often in agriculture, an experimenter is interested in comparing the yielding abilities of several varieties of a crop, say wheat. Similarly, there may be four different drugs for the control of blood pressure and it is of interest to know whether these four drugs are equally efficient in the control of blood pressure.

In each of the above examples, we have several populations (one for each method of teaching, each crop variety, each drug) and the hypothesis that is required to be tested is whether the means of these populations are equal. Analysis of variance, written in short as ANOVA is useful in such situations. We shall assume that each of the populations have a normal distribution with possibly different means but having the same variance.

We shall start this unit by acquainting you with real life problem in which you need to compare means of several populations simultaneously. We shall introduce the method of analysis of variance through this example. In this unit we shall confine our attention to study the effect of a single factor on a variable under study. Such study lead to one-way classification data. We shall then formulate the model for one-way classification for the problem considered and use this model to explain the procedure to

have better insight about the model parameters we have also discussed the estimation of model parameters and their pairwise comparison. We shall end our discussion in this unit by acquainting you with a concept of random effects.

Objectives

After reading this unit, you should be able to

- identify different sources of variation in a given problem;
- distinguish the one-way classification data from other types;
- write down the model for one-way classification problems;
- carry out the test of hypothesis on equality of all treatment means, equality of pairs of treatments means;
- estimate the model parameters and provide confidence intervals;
- draw inferences and conclusions from the analysis.

8.2 ANALYSIS OF VARIANCE : BASIC CONCEPTS

We shall start with a simple real life problem that many of us face. Now a days most of us use gas for cooking purposes. Most of the gas users are customers of big companies. The customers get their refills (filled gas cylinders) through the agents of these companies. One of the customers, Mrs. Devi, who is attached to ABC agency, has faced a problem in the recent past. She observed that her cylinders were not lasting as long as they used to be in the past. So she suspected that the amount of gas in the refills was less compared to what she used to get in the past. She knew that she is supposed to get 14.2 kgs of gas in every refill. She explained her problem to the customers' redressal cell of the company.

Subsequently, the company made a surprise check on an ABC agent. They took 25 cylinders that were being supplied to customers from this agency and measured the amount of gas in each of these cylinders. The 25 observations were statistically analysed and through a simple test of hypothesis (you may recall how this is done) it was inferred that the mean amount of gas in the cylinders supplied by ABC agency was significantly lower than 14.2 kgs. On investigation, it was revealed that the agent was tapping gas from cylinders before they are being supplied to the customers.

There were five agents of the company in the town where Mrs. Devi was living. To protect customers' interests, the company decided to carry out surprise checks on all the agents from time to time. During each check, they picked up 7 cylinders at random from each of the five agents resulting in the data given in Table-1. Is it possible to test from this data whether the mean amount of gas per cylinder differs from agent to agent? Well, it is possible to carry out a simple test of hypothesis for each of the agents separately. But there is a better statistical procedure to do this simultaneously. We shall see how this can be done.

-
- E1) What is the difference between the hypothesis you studied in Unit 6 and the hypothesis in the above problem?
-

Before we proceed further with the above problem, we introduce to you some concepts and terminology.

You know that variation is inevitable in almost all the variables (measurable characteristics) that we come across in practice. For example, the amount of gas in two refills is not the same irrespective of whether the gas is tapped or not. Consider the data in Table-1.

Table-1: Data on Gas Weights (kgs)

S. No.	Agent				
	1	2	3	4	5
1	14.36	13.51	13.74	14.09	13.51
2	14.15	13.60	13.94	14.52	14.03
3	14.20	13.71	14.10	14.58	13.84
4	14.12	13.94	13.93	14.08	13.29
5	14.05	13.95	13.61	13.66	13.97
6	14.15	13.67	13.97	13.90	13.44
7	14.17	13.79	13.88	14.13	13.94

We have the weights of gas in 35 cylinders taken at random, seven from each of the five agents. These 35 weights exhibit variation. You will agree that some of the possible reasons for this variation are one or more of the following:-

- The gas refilling machine at the company does not fill every cylinder with exactly same amount of gas.
- There may be some leakage problem in some of the cylinders.
- The agency/agents might have tapped gas from some of these cylinders.
- All the 35 cylinders are not filled by the same filling machine.

Thus, the variation in the 35 weights might have come from different sources. Though the variation is attributable to several sources, depending upon the situation, we will be interested in analysing whether most of this variation can be due to differences in one (or more) of the sources. For instance, in the above example, the company will be interested in identifying if there are any differences among the agents. So the **source of variation** of interest here is **agents**. In other words, we are interested in one factor or, one-way analysis of variance. Before continuing further with the discussion, try to identify the sources of variations in the following exercises.

-
- E2) Five different fertilisers were tested on a particular crop to compare their performances. Each fertiliser was applied on 4 plots. All the twenty plots used here belong to the same location. Identify the sources of variation and mention which of these you suspect to have larger contribution to the variation.
- E3) One of several components manufactured by an automobile company is the drive shaft. This component is manufactured on a special purpose machine which has 8 stations. Drive shafts are produced simultaneously on all the eight stations. Milling is one of the operations carried out at each station. Milling operation produces a slot on the drive shaft and the length of the slot, x , should be between 5.85 cms and 6.05 cms. Identify possible sources that contribute to variation in x .
-

8.2.2 One-Way Classification

Now that you know what is source of variation, you can think of different types of sources. In the gas company example, agents form one type of source. If the cylinders

under consideration were refilled by different filling machines, then filling machines is another type of source of variation. Similarly, in E2), supposing that 10 of the 20 plots belong to one location and the remaining 10 belong to another location, then there are two types of sources of variation, namely, 'fertilisers' and 'location'. When the data are classified only with respect to one type of source of variation, we say that we have one-way classification data. On the other hand, in the above fertiliser example, if we also have reasons to believe that there could be differences between locations, then we have a two-way classification data. In many situations, one conducts experiments to study the effect of a single factor on a variable under study. Such experiments, known as one-factor experiments, lead to one-way classification data. The following is an example of one such experiment.

Example 1: The tensile strength of synthetic fibre used to make cloth is of interest to the manufacturer. It is suspected that the strength is affected by the percentage of cotton in the fibre. The cloth was produced by varying the cotton percentage at five different levels, namely, 15%, 20%, 25%, 30% and 35%. Samples were drawn from the cloth produced at each of these levels and the tensile strengths were measured. In this way one-way classification data was obtained.

* * *

As we have already mentioned, in this unit, we will confine ourselves to one-way classification only. We shall now look at the theoretical model for the one-way classification. To make it easy for you to understand, we shall describe the model with the gas company example.

8.3 MODEL FOR ONE-WAY CLASSIFICATION

As you already know the first step in any modelling process is to identify the various parameters involved in the problem. We shall now specify these parameters for the problem under consideration. Consider the problem of the gas company example. It is suspected that the agents might tap the gas from each of the cylinders. Let us say that agent i , taps a_i kgs of gas from each of the cylinders supplied by him, $i = 1, 2, \dots, k$. Here k is the number of agents. Recall that $k = 5$ for Mrs. Devi's town. We shall assume that the weight of each cylinder supplied by the company to the agents varies marginally around 14.2 kgs, so that the average weight of a cylinder is given by $\mu = 14.2$ kgs. Next, let us say that, according to the vigilance policy, the company picks up n_i cylinders at random from agent i and measures the weights of the gas. For the data in Table-1, n_i s are all equal to 7. Let y_{ij} be the weight of the gas in j^{th} cylinder from agent i . Then, we would expect the value of y_{ij} to be 14.2 kgs provided there is no tapping of gas. However, since agent i taps a_i kgs of gas, we would expect y_{ij} to be $14.2 - a_i$ kgs. and this should be true for $j = 1, 2, \dots, n_i$. Yet, when we measure y_{ij} , it would not be exactly equal to $14.2 - a_i$ kgs. There could be several reasons for this. When the cylinder was supplied to the agent, its weight might be slightly less than or more than 14.2 kgs. There could have been minor variation in reading the measurement. Due to these reasons, it is reasonable to assume that $y_{ij} = \mu - a_i + \epsilon_{ij}$, where ϵ_{ij} , called the error, is the difference between the observed weight and the true weight. Letting $\tau_i = -a_i$, for $1 \leq i \leq k$, we can write

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad j = 1, 2, \dots, n_i, \quad i = 1, 2, \dots, k. \quad (1)$$

In the above equation, μ is called the general (mean) effect and τ_i is called the effect of agent i . In general, μ and τ_i s are unknown quantities but are assumed to be constants. These are known as the model parameters. On the other hand, the ϵ_{ij} s are due to random fluctuations and hence are assumed to be random variables. As a result, y_{ij} s are also random variables.

Stating The Hypothesis

If all the agents are honest, then τ_i s must all be equal to zero. Thus, the company would be interested in testing the hypothesis

$$H_0: \tau_1 = \tau_2 = \dots = \tau_k = 0. \quad (2)$$

The alternative hypothesis here is that at least one of the τ_i s is different from zero (or equivalently, $\tau_l \neq \tau_m$ for some l and m).

In Model (1) above, we make the following two assumptions:

- i) The ϵ_{ij} s are independent and identically distributed random variables with mean 0 and variance σ^2 .
- ii) ϵ_{ij} s follow normal distribution.

The model is a typical example of a **model for one-way classification**.

In Eqn.(2), H_0 states that all τ_i s are equal to zero. However, in general, we may not always be interested in testing the equality of means to zero. One may just want to test the equality of means only, that is, want to test the hypothesis

$$K_0 : \tau_1 = \tau_2 = \dots = \tau_k$$

Fortunately, it can be shown that the test procedure is same for both H_0 and K_0 .

On the same lines as above you may now try to construct the linear model for the fertiliser example.

- E4) Consider E2). Assuming that all the twenty plots are homogeneous, construct a linear model for this problem and explain various quantities you define.

In the following section, we will learn the analysis of variance (ANOVA) technique to carry out the tests of hypothesis of the type mentioned in Eqn.(2) above.

8.4 TEST PROCEDURE

As the name suggests, in ANOVA, we analyse the variability in any given set of data and try to see if this variability is mainly due to certain specific reasons. We first try to understand the procedure through the gas company example.

8.4.1 Sums of Squares

Consider the gas company example and the model given in Eqn. (1). Look at seven weights under any agent in Table-1. You find there is variation. This variation arises from several sources such as measurement error etc. But these sources are common to all the agents. So it is reasonable to expect the variance of the observations under each agent is the same for all agents. If this common variance is represented by σ^2 , then to test the null hypothesis that the k population means are all equal, we shall compare two estimates of σ^2 — one based on the variation among the sample means, and one based on the variation within the samples.

A pooled estimate of σ^2 i.e., the mean of the sample variances s_i^2 , is given by

$$\sigma^2 = \frac{\sum_{i=1}^5 s_i^2}{5} = \frac{\sum_{i=1}^5 \sum_{j=1}^7 (y_{ij} - \bar{y}_i)^2 / 6}{5} = \frac{1.5306}{30} = 0.05102$$

The quantity $\sum_{i=1}^5 \sum_{j=1}^7 (y_{ij} - \bar{y}_i)^2$ is called the **error sum of squares (SS_e)**, the quantity that estimates random (or chance) error.

Now look at the agent wise averages \bar{y}_i s in Table-2. If the agents are uniform, we expect the variance of \bar{y}_i s to be small. The sample variance of \bar{y}_i s is given by

$$\frac{\sum_{i=1}^5 (\bar{y}_i - \bar{y})^2}{4} = \frac{1.2976}{4} = 0.3244$$

where \bar{y} is the average of the \bar{y}_i s. The quantity $\sum_{i=1}^5 (\bar{y}_i - \bar{y})^2$ is called the **sum of squares due to agents (SS_a)**.

The sample variance of all the 35 weights is given by

$$\frac{\sum_{i=1}^5 \sum_{j=1}^7 (y_{ij} - \bar{y})^2}{34} = \frac{2.8282}{34}$$

The quantity $\sum_{i=1}^5 \sum_{j=1}^7 (y_{ij} - \bar{y})^2$ is called the **total sum of squares (TSS)**.

Observe that $2.8282 = 1.2976 + 1.5306$. That is, $TSS = SS_a + SS_e$. Therefore, if SS_a is large compared to SS_e , we can infer that τ_i s are indeed different. However, the term 'large' or 'small' are subjective. In order to quantify the largeness of SS_a related to SS_e , we compute the ratio

$$F = \frac{SS_a/4}{SS_e/30} = 6.35.$$

This ratio, under the hypothesis, is known to have an F-distribution on 4 and 30 degrees of freedom. Comparing this ratio with the table value of F-distribution with 4 and 30 degrees of freedom at say, 5 % level of significance allows one to decide whether to reject the hypothesis of equality of means. We shall talk more about it later. Let us now consider a general case. Throughout this unit we shall denote the total number of observations by N, i.e., $N = \sum_{i=1}^k n_i$. Let these N observations be

$$y_{11}, y_{12}, \dots, y_{1n_1}, y_{21}, y_{22}, \dots, y_{2n_2}, \dots, y_{k1}, y_{k2}, \dots, y_{kn_k}.$$

The **total sum of squares (TSS)**, of all these observations is given by

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2,$$

where \bar{y} is the average of all the y_{ij} s. Let \bar{y}_i be the average weight of the gas in cylinders from agent i , i.e., $\bar{y}_i = (\sum_{j=1}^{n_i} y_{ij})/n_i$. Using simple algebra it can be shown that

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 \quad (3)$$

We are leaving this for you to verify yourself.

E5) Show that $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$.

Using $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, we see that

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (\epsilon_{ij} - \bar{\epsilon}_i)^2,$$

where $\bar{\epsilon}_i = (\sum_{j=1}^{n_i} \epsilon_{ij})/n_i$.

You may note here that $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ involves only the errors ϵ_{ij} s. We call this sum of **squares due to error** and denote it by SS_e . Similarly, $\sum_{i=1}^k n_i(\bar{y}_i - \bar{y})^2$ indicates the dispersion among the agents' averages. This is called the **sum of squares due to agents** (or **sum of squares due to treatments** in general) and we denote this by SS_a (SS_{tr} in general).

When all the n_i s are equal, say equal to n , the data are said to be **balanced**. In our further discussion we shall assume that the data are balanced, unless mentioned otherwise. Thus we have from Eqn.(3) above

$$TSS = SS_{tr} + SS_e.$$

i.e., the total sum of squares is decomposed into sum of squares due to error and sums of squares due to treatments. When all the agents are honest, we expect all the \bar{y}_i s to be close to \bar{y} and hence SS_a to be small. Under Assumption (i) ϵ_{ij} s are independent and identically distributed (iid) random variables with mean 0 and variance σ^2 , it can be shown that the statistical expectation of $SS_{tr}/(k - 1)$, called the **mean sum of squares due to treatments**, and denoted by MS_{tr} , is given by

$$E(MS_{tr}) = \sigma^2 + \frac{\sum_{i=1}^k n_i \tau_i^2}{(k - 1)}. \quad (4)$$

Also, as per Assumption (ii) above ϵ_{ij} s are normally distributed, then under H_0 , SS_{tr}/σ^2 follows χ^2 -distribution with $(k - 1)$ degrees of freedom (df).

On the other hand, the mean squares for error, $SS_e/(nk - k)$ denoted by MS_e , is a measure of natural variability present in the data due to non-assignable causes and is an unbiased estimator of σ^2 .

Thus, under Assumptions (i) and (ii), it follows that SS_e/σ^2 has χ^2 -distribution with $\nu = (nk - k)N - k$ df. In fact, MS_e acts as yardstick to decide whether SS_{tr} is large or small. To decide whether to reject H_0 or not, we need to compute the ratio

$$F_0 = \frac{MS_{tr}}{MS_e} = \frac{SS_{tr}/(k - 1)}{SS_e/\nu}.$$

This ratio follows **F-distribution** with **degrees of freedom $k - 1$ and ν** . Recall that type I error, α , is the chance of rejecting H_0 when it is true.

Suppose we decide to carry out the test at (5% level of significance), all that we have to do is to get the tabulated F-value with respective df from the F-table(ref. Table-1 in the Appendix given at the end of the block.) and reject H_0 if F_0 is greater than this tabulated F-value.

Let us now apply this test to our gas company example. We first prepare the ANOVA table for gas company example.

Recall the sampling distributions you have studied in Unit 4.

8.4.2 Preparation of ANOVA Table

Consider the data in Table-1. Let us first summarise these data by computing the agent-wise totals and averages. Let y_i denote the total of y_{ij} s under agent i . Then $\bar{y}_i = (\sum_{j=1}^7 y_{ij})/7$, $i = 1, 2, \dots, 5$. These values are summarised in Table-2.

Table-2: Summary of Data on Gas Weights

Statistic	Agent				
	1	2	3	4	5
Total (y_i)	99.20	96.17	97.17	98.96	96.02
Average (\bar{y}_i)	14.1714	13.7386	13.8814	14.1371	13.7171
Sample variance, $s_i^2 = \sum_{j=1}^7 \frac{(y_{ij} - \bar{y}_i)^2}{6}$	0.00911	0.02748	0.02585	0.10536	0.08812
$6 \times s_i^2$	0.05469	0.16489	0.15509	0.63214	0.52874

To carry out the test of hypothesis H_0 given in Eqn. (1), we prepare a table called the ANOVA table. For the one-way classification, the ANOVA Table typically looks like Table-3.

Table-3: Model ANOVA Table For One-Way Classification

SV	DF	SS	MS	F-Ratio
(source of variation)	(degrees of freedom)	(sum of squares)	(mean squares)	
Treatments	$k - 1$	SS_{tr}	$MS_{tr} = \frac{SS_{tr}}{k-1}$	$\frac{MS_{tr}}{MS_e}$
Error	$N - k$	SS_e	$MS_e = \frac{SS_e}{k(n-1)}$	
Total	$N - 1$	TSS		

To prepare the ANOVA table for the gas company example, we need to compute TSS, SS_a and SS_e . We shall now give the formulas for computation of these quantities for the general (unbalanced) case.

$$TSS = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - CF, \quad (5)$$

$$SS_{tr} = \sum_{i=1}^k \frac{y_i^2}{n_i} - CF, \quad (6)$$

$$\text{and } SS_e = TSS - SS_{tr}, \quad (7)$$

where CF is the correction factor and is given by $CF = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2}{N} = N\bar{y}^2$ and n_i s are as defined in model (1). For the problem in question which is a balanced problem with $n = n_i$ for $i = 1, 2, \dots, 7$, we have

$$\bar{y} = 13.9291, \quad CF = 6790.7357,$$

$$\sum \sum y_{ij}^2 = 6793.5698, \quad TSS = 2.8341,$$

$$SS_{tr} = \frac{(99.20)^2 + (96.17)^2 + (97.17)^2 + (98.96)^2 + (96.02)^2}{7} - CF \\ = 6792.0343 - 6790.7357 = 1.2986,$$

and

$$SS_e = TSS - SS_{tr} = 2.8341 - 1.2986 = 1.5355$$

The ANOVA table for the problem is given in Table-4 below.

Table-4: ANOVA For GC Example Data

SV	DF	SS	MS	F-Ratio
Agents	4	1.2986	0.3247	6.35
Error	30	1.5355	0.0511	
Total	34	2.8341		

The tabulated value of P with 4 and 30 degrees of freedom at 5% level of significance is equal to 2.69. Since the calculated F-value from the ANOVA table is larger than the

tabulated value, we reject the null hypothesis H_0 , inferring thereby that the agents indeed do differ in respect of average weight of gas cylinder. In the case of balanced data above calculation can be simplified by using the special tips given in the following remark.

Remark: In case of balanced data if you compute the standard deviation, then you can compute the sum of squares TSS and SS_{tr} directly without computing the correction factor, CF. To get TSS, input all the y_{ij} s and get their standard deviation (SD). Then TSS is equal to square of this SD multiplied by the number of observations. For the above problem, this standard deviation is equal to 0.28426 and

$TSS = 35 \times (0.28426)^2 = 2.8281$. To get SS_{tr} , compute the SD of the totals (y_i s), square it and then multiply it by $\frac{k}{n}$. For the above problem, SD of the totals is equal to 1.3483, so that

$SS_{tr} = (1.3483)^2 \times \frac{5}{7} = 1.2985$ (the discrepancy is due to rounding of errors). Next, SS_e is obtained by subtraction. Thus, $SS_e = 2.8281 - 1.2985 = 1.5296$.

Remember these special tips apply only if the data are balanced, i.e., all the n_i s are equal. Do not apply this to unbalanced data.

Before we continue with gas company example, you must get some practice on preparing the ANOVA table and carrying out the test of hypothesis for equality of means. You can do that while trying the following exercises.

- E6) Continuing with E(3), it was reported that 5% of the drive shafts were being rejected due to non conformance of slot length, x , to its specifications. In order to diagnose the problem, ten drive shafts were collected from each of the 8 stations and their lengths were measured. These data are presented in Table 5. To reduce your efforts on computation some summary statistics are also incorporated in the table. If the settings of the special purpose machine were proper, the mean shaft lengths should be same for all the 8 stations. Now do the following.
- Frame the null hypothesis for equality of all means by introducing the necessary notation.
 - State the alternative hypothesis .
 - Prepare the ANOVA table and carry out the test of hypothesis.

Table 5: Slot Lengths (At the Investigation Stage)

S. No.	Station							
	1	2	3	4	5	6	7	8
1	5.85	5.92	5.87	6.01	6.02	5.87	5.94	6.02
2	5.89	5.95	5.91	5.95	5.92	5.90	5.91	6.07
3	5.90	5.91	5.94	5.90	6.00	5.92	5.95	6.00
4	5.92	5.88	5.92	5.93	5.94	5.95	5.96	6.03
5	5.95	5.92	5.98	5.97	5.97	5.96	5.98	6.11
6	5.91	5.92	5.92	5.94	5.98	5.94	5.95	6.00
7	5.88	5.95	5.90	5.96	5.96	5.92	5.88	5.99
8	5.95	5.90	5.89	5.98	6.00	5.91	5.93	6.01
9	5.92	5.93	5.95	5.89	5.94	5.95	5.97	5.98
10	5.84	5.94	5.95	5.97	5.97	5.98	5.98	6.02
Total	59.01	59.22	59.23	59.50	59.70	59.30	59.45	60.23
Average	5.901	5.922	5.923	5.950	5.970	5.930	5.945	6.023
$\sum_{j=1}^{10} y_{ij}^2$	348.2305	350.7052	350.8289	354.037	356.4178	351.6584	353.4393	362.7793

- E7) After analysing the data in E6), it was found that the 8 stations were not uniform. Investigation by the concerned personnel revealed that there were discrepancies in the settings of some fixtures used in the stations. So the settings were adjusted and data on the slot lengths were collected again. These data are presented in Table-6.

Here SD is the population SD, not the sample SD.

This study was carried out in a leading automobile industry in Bangalore.

Problem was rectified and the rejections due to slot length non conformance were eliminated.

Table-6: Slot Lengths (After Corrective Action)

S. No.	Station							
	1	2	3	4	5	6	7	8
1	5.87	5.86	5.97	5.94	5.88	5.93	5.88	5.87
2	5.94	5.93	5.88	5.92	5.93	5.90	5.90	5.92
3	5.92	5.96	5.92	5.90	5.92	5.96	5.92	5.92
4	5.91	5.86	5.89	5.95	5.95	5.95	5.98	5.89
5	5.89	5.92	5.95	5.91	5.91	5.92	5.92	5.95
6	5.95	5.94	5.91	5.91	5.96	5.94	5.93	5.94
7	5.88	5.86	5.94	5.98	5.89	5.92	5.85	5.90
8	5.91	5.88	5.84	5.88	5.92	5.97	5.91	5.87
9	5.97	5.90	5.94	5.93	5.95	5.98	5.94	5.92
10	5.86	5.92	5.92	5.97	5.95	5.95	5.93	5.93
Total	59.10	59.03	59.16	59.29	59.26	59.42	59.16	59.11
Average	5.910	5.903	5.916	5.929	5.926	5.942	5.916	5.911
$\sum_{j=1}^{10} y_{ij}^2$	349.29	348.47	350.01	351.53	351.18	353.08	350.01	349.41

Do (a), (b) and (c) as given in E6).

In the next section we shall consider comparison of pairs of treatment means.

8.5 PAIRWISE COMPARISONS

Having rejected the hypothesis of equality of means, one might wish to compare pairs of treatment means. Though there are several methods to do this, we shall confine our attention to the **least significant difference (LSD)** method.

Supposing we want to compare the treatment means τ_1 and τ_m . An unbiased estimator of $\tau_1 - \tau_m$ is $\bar{y}_1 - \bar{y}_m$ which has variance $(\frac{1}{n_1} + \frac{1}{n_m})\sigma^2$. Furthermore, under the normality assumptions, $\bar{y}_1 - \bar{y}_m$ is normally distributed and is independent of MS_e . Consequently, under $H_0 : \tau_1 = \tau_m = 0$,

$$t_0 = (\bar{y}_1 - \bar{y}_m) / \sqrt{MS_e / (\frac{1}{n_1} + \frac{1}{n_m})} \quad (8)$$

follows t-distribution with ν df where ν is the error df as before. The hypothesis H_0 is rejected if the absolute value of $(\bar{y}_1 - \bar{y}_m)$ is larger than the

$$LSD = t_{\alpha/2, \nu} \times \sqrt{MS_e / (\frac{1}{n_1} + \frac{1}{n_m})}.$$

You may recall that in the gas company example, we have rejected the hypothesis that the treatment means are equal. Looking at the averages in Table-2, we have reasons to suspect that agents 2,3 and 5 have possibly different means. Before investigating these agents, let us see if the other two agents 1 and 4, have the same mean. First let us examine whether $\tau_1 = \tau_4$. Since all n_i s are equal to 7 in this example, we have $LSD = 1.96 \times \sqrt{2 \times 0.051/7} = 0.2366$. Since $\bar{y}_1 - \bar{y}_4 = 14.1714 - 14.1371 = 0.0343$ is less than LSD, we cannot reject the hypothesis that $H_0 : \tau_1 = \tau_4$. We therefore conclude that there is no substantial evidence to suspect that agents 1 and 4 do differ in respect of their means.

It is now time for you to try the following exercises to make sure that you understand what is going on.

- E8) Consider the data in Table-5. Frame the hypothesis (ie., write down H_0 and H_1) for each of the following cases and carry out the tests at 5% level of significance.
- a) Test whether there is any difference between stations 1 and 6 with respect to mean slot length

- b) Do the similar exercise as in (a) for stations 4 and 6.
- c) Assuming that there is no difference between stations 4 and 7, test whether the common mean slot length for these two stations is equal to the target (since the specifications are 5.85 to 6.05, the target is the middle point which is 5.95).
- d) Test whether the mean slot length for station 3 is significantly lower than the target.
- E9) Consider the data in Table-6. Assuming that mean slot length is same for all the 8 stations, test whether slot length is set at the target.
-

Let us resume our analysis of gas company example and examine the agents 2, 3 and 5. Arranging the 5 averages in the descending order and computing the differences, we get

$$\bar{y}_1 = 14.1714$$

$$\bar{y}_4 = 14.1371 \quad \bar{y}_1 - \bar{y}_4 = 0.0343$$

$$\bar{y}_3 = 13.8814 \quad \bar{y}_4 - \bar{y}_3 = 0.2557$$

$$\bar{y}_2 = 13.7386 \quad \bar{y}_3 - \bar{y}_2 = 0.1428$$

$$\bar{y}_5 = 13.7171 \quad \bar{y}_2 - \bar{y}_5 = 0.0215.$$

Since $LSD = 0.2366$, we find that there is significant difference between τ_4 and τ_3 .

However, pairwise comparisons among τ_2 , τ_3 and τ_5 do not show up significant differences. From these analyses we can infer that customers get a bad deal from agents 2,3, and 5.

In many situations, the data are such that the number of observations under a treatment is not the same for all treatments. This might happen due to constraints or due to accidents. For instance, in the fertiliser example (E2), if we only had 18 plots available for the study, then it would not be possible to try the five fertilisers on the same number of plots. Two of the fertilisers may have to be applied on 3 plots each, while the other 3 on four plots each. Or it might be that 20 plots were available for the study and fertilisers were applied on the same number of plots, but due to unforeseen conditions, crops in two of the plots got damaged. In the next section, we give details of the analysis for unbalanced data.

8.6 UNBALANCED DATA

In the unbalanced case, the sum of squares for the ANOVA table can be computed using the formula given by (5), (6), and (7). We shall now work out an example to point out the minor differences in the analysis.

Example 2: An experiment was conducted to determine whether four specific firing temperatures affect the density of a certain type of brick. The data are presented in Table-7 along with some summary statistics.

Table-7: Density of Bricks and Summary Statistics

		Temperature			
		100°C	125°C	150°C	175°C
Density	21.8	21.7	21.9	21.9	
	21.9	21.4	21.8	21.7	
	21.7	21.5	21.8	21.8	
	21.6	21.4	21.6	21.4	
	21.7	-	21.5	-	
n_i		5	4	5	4
Total, y_i		108.7	86.0	108.6	86.8
\bar{y}_i		21.74	21.50	21.72	21.7
$\sum_{j=1}^n y_{ij}^2$		2363.19	1849.06	2358.90	1883.70

Let τ_1, τ_2, τ_3 and τ_4 be the effects of temperature on density at 100°C, 125°C, 150°C and 175°C respectively. Let μ be the general effect. To test the hypothesis, $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4$, we compute the sum of squares and form the ANOVA Table. We have

$$\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij} = 108.7 + 86.0 + 108.6 + 86.8 = 390.1, \sum n_i = 18, \bar{y} = \frac{390.1}{18} = 21.6722,$$

$$\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij}^2 = 2363.19 + 1849.06 + 2358.90 + 1883.7 = 8454.85,$$

$$CF = \frac{(390.1)^2}{18} = 8454.3338, TSS = 8454.85 - 8454.3338 = 0.5161,$$

$$SS_a = \sum_{i=1}^4 \frac{(y_i)^2}{n_i} - CF = \frac{(108.7)^2}{5} + \frac{(86.0)^2}{4} + \frac{(108.6)^2}{5} + \frac{(86.8)^2}{4} - 8454.3338 = 0.1562,$$

and $SS_e = TSS - SS_a = 0.5161 - 0.1562 = 0.3599$. We can now write down the ANOVA Table.

Table-8: ANOVA for the Density Data

SV	DF	SS	MS	F
Temperature	3	0.1562	0.0520	2.02
Error	14	0.3599	0.0257	
Total	17	0.5161		

Remember that the df for error is always equal to total number of observations minus the number of treatments.

Since the observed F-ratio, 2.02, is less than the tabulated F-value at 5 % confidence level with 3 and 14 df which is 3.34, we cannot reject the null hypothesis.

Let us remember that not being able to reject H_0 does not mean that the treatment means (τ_i s) are equal. Let us test if τ_1 and τ_2 are significantly different. In other words, we wish to test

$$H_0 : \tau_1 = \tau_2 \quad Vs \quad H_1 : \tau_1 \neq \tau_2.$$

We shall now compute t_0 given by Eqn.(8) which has t-distribution with 14 df. Substituting the values, we get $t_0 = 2.231$, whereas the tabulated $t_{0.025, 14}$ is equal to 2.145 (See Table-2 in the Appendix given at the end of the block.). Hence, we conclude that there is significant difference (at 5% level of significance) between τ_1 and τ_2 .

* * *

You may now try the following exercises to reinforce your learning for the unbalanced case.

E10) In the above example, test if τ_2 and τ_3 are significantly different (at 5% level).

E11) A tailoring shop owner has 3 tailors, A, B and C, working under him who stitch only men's shirts. During a particular week, the owner tried to study their efficiencies with regard to productivity and obtained the following data.

Table-9 : Tailors' Productivity Data

Day	Number of shirts stitched per shift		
	A	B	C
1	6	7	9
2	8	8	8
3	6	7	*
4	5	5	6
5	*	8	*

* Tailor was on leave.

Prepare a summary table, present the ANOVA table and test (at 5% level) whether all the 3 tailors are equally productive.

In the following section, we will acquaint you with a concept known as random effects. Our purpose here is not give you the details but to make you understand broadly the differences between fixed effects and random effects. Again here, we shall concentrate on the balanced data.

8.7 RANDOM EFFECTS MODEL

In all the examples that we have seen so far in this unit, the treatment levels were all fixed. In the gas company example, the agents were five specific agents. They are not picked up at random from a group of agents. Similarly, in the fertiliser example, we were interested in the effects of five specific fertilisers. Since the treatment levels are fixed, we assumed that their effects, τ_i s, are also fixed constants. Therefore, these effects are called **fixed effects** and in this case the model (1) is called a **fixed effects model**.

In some situations, the levels of treatment are picked up at random and then the data are collected at these random levels. As an example, let us once again look at the gas company example, but this time we will confine ourselves to the factory where the cylinders are refilled. One such factory has 500 filling stations. Cylinders are refilled simultaneously at each of these stations.

One of the things that the quality control engineer has to ensure is that the stations are homogeneous. That is, he must ensure that the mean amount of gas filled by a station is the same for all the stations. To do this, the engineer selects, from time to time, 5 stations at random, and from each station he picks up five cylinders, again at random, and measures the amount of gas in each of these cylinders. One such set of data is presented in Table-10.

Table-10 : Gas Weights Data

S.No.	Station Number				
	5	47	193	301	398
1	14.24	13.92	14.16	14.19	14.06
2	14.24	13.95	14.18	14.20	14.05
3	14.26	13.96	14.17	14.19	14.06
4	14.25	13.97	14.14	14.19	14.03
5	14.26	13.95	14.15	14.21	14.06
$\sum y_i$	71.25	69.75	70.8	70.98	70.26
$\sum \bar{y}_i$	14.25	13.95	14.16	14.2	14.05
$\sum y_{ij}^2$	1015.3129	973.0139	1002.529	1007.6324	987.2942

Let τ_i be the effect of station i on the gas weight. Since stations are selected at random, we may assume that τ_i s are random variables. Note that the data are one-way classification data. We use the model given in Eqn.(1) in this case also but interpret the τ_i s as random variables. In fact, we assume that the τ_i s are iid with mean 0 and variance σ_τ^2 . The ϵ_{ij} s play the same role as in the fixed effects model. The new model with τ_i s as random variables is called the **random effects model**.

The engineer tries to test the homogeneity of all the 500 stations. This is done by testing

$$H_0 : \sigma_\tau^2 = 0 \quad \text{vs} \quad H_1 : \sigma_\tau^2 > 0.$$

The test procedure is exactly the same as the one for the fixed effects model. That is, prepare the ANOVA table and check if the observed F-value is larger than the tabulated F-value. If this is the case, then reject H_0 and infer that the stations are not homogeneous.

In random effects model, σ^2 and σ_τ^2 are called the **variance components**. In case of the **balanced data**, they can be estimated unbiasedly by

$$\hat{\sigma}^2 = MS_e \quad \text{and} \quad \hat{\sigma}_\tau^2 = \frac{MS_{tr} - MS_e}{n}.$$

Now let us analyse the data of Table-10.

We have

$$\sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij} = 353.04, \sum n_i = 25, \bar{y} = 14.1216$$

$$\sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij}^2 = 4985.7824$$

$$C.F. = \frac{(353.04)^2}{25} = 4985.4897$$

$$TSS = 0.2927$$

$$SSa = \sum \frac{y_i^2}{n_i} - CF \\ = 4985.7786 - 4985.4897 \\ = 0.2889$$

$$SSe = TSS - SSa \\ = 0.0038$$

The ANOVA table is given below.

Table-11: ANOVA For Gas Weights Data

SV	DF	SS	MS	F
Stations	4	0.2889	0.07223	380.16
Error	20	0.00380	0.00019	
Total	24	0.29273		

We see that the observed F-value is very large while the tabulated F-value at 5 % level of significance is equal to 2.87. If you observe the data in Table-10, you find there is some problem with station 47. Probably there was some setting problem or something else. Since the 5 stations are selected at random from the 500 stations, we conclude that there may be problems with other stations as well. So this calls for further investigation.

The estimates of σ^2 and σ_τ^2 are

$$\hat{\sigma}^2 = 0.00019 \quad \text{and} \quad \hat{\sigma}_\tau^2 = \frac{0.07223 - 0.00019}{5} = 0.0144.$$

The engineer's aim must be to reduce σ_τ^2 as much as possible by improving the uniformity among the stations. This might perhaps be achieved by setting the faulty stations right.

And now an exercise for you.

- E12) In textile mill there are 300 looms. It is known that the performance of the looms affect the strength of the fabric. Four looms were selected at random, and three samples were tested from the fabric produced from each of these looms. Test if the looms in the company are homogeneous. Estimate the variance components.

Table-12 : Fabric Strength Data

Loom	Observations				Average
1	98	98	99	96	97.75
2	91	90	93	92	91.50
3	96	95	97	95	95.75

We now end this unit by giving a summary of what we have covered in it.

8.8 SUMMARY

In this unit we have learnt the following important points.

- 1) In real life situations, we often encounter problems in which we need to study several populations.
- 2) Analysis of variance is a useful technique to test the equality of means of several populations and carry out subsequent analysis.
- 3) One-way classification data arise when we are interested in studying the effect of a single source of variation on a variable.
- 4) Building a linear model for the one-way classification data.
- 5) Preparation of the ANOVA table and testing the equality of all treatment means.
- 6) Estimation of model parameters.
- 7) The least significant difference method for comparison of all pairs of treatment means.
- 8) Analysis of the unbalanced one-way classification data.
- 9) The concept of random effects model.

8.9 SOLUTIONS/ANSWERS

- E1) In Unit 6, the tests of hypotheses were concerning either one population mean being equal to a constant or equality of two population means. That is, they were of the type $H_0 : \mu = 14.2$ or $\tau_1 = \tau_2$. In gas company example, the hypothesis is concerning five populations, that is, $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5$.
- E2) Here fertilisers will be the main source of variation of interest. If the fertility of the soil of the 20 plots varies widely, then the plots should also be considered as a major source of variation. However, since it is given that all the twenty plots belong to the same location, it is expected that their effect on the yield is same. In this case, the plots may not be treated as a major source of variation.
- E3) The main source of variation of interest here is 'stations'. We should be, in the first place, interested in examining if the output from all the eight stations is uniform. Besides stations, if the machine is operated by different operators at different times, then the operators could be another source of variation.
- E4) Let y_{ij} be the crop yield of j^{th} plot on which fertiliser i is applied. The general mean μ may be interpreted as the average yield per plot if no fertiliser is applied. Let τ_i be average additional yield per plot if fertiliser i is applied. Then the

expected value of y_{ij} is $\mu + \tau_i$. Since y_{ij} s are subjected to random fluctuations, we can write $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $j = 1, 2, \dots, 4$, $i = 1, 2, \dots, 5$, where ϵ_{ij} s are the deviations from the expected means. The variation in ϵ_{ij} s is due to several causes that are common to all yields. For these reasons, ϵ_{ij} s are assumed to have mean zero and variance σ^2 . Therefore, the linear model is given by

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad j = 1, 2, \dots, 4, \quad i = 1, 2, \dots, 5,$$

where ϵ_{ij} s are iid random variables with mean 0 and variance σ^2 .

E5) We can write

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$$

squaring both sides and summing on i and j, we have

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 \\ &\quad + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) \dots \end{aligned} \quad (9)$$

Now

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) = \sum_{i=1}^k (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$$

since \bar{y}_i is the mean of the ith sample, and hence $\sum_{j=1}^n (y_{ij} - \bar{y}_i) = 0 \forall i$

Also, in (Eqn.9) above

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

E6) (a) Let μ be the general mean and let τ_i be the effect of station i on x. Let x_{ij} be the slot length of j^{th} drive shaft from i^{th} station, $j = 1, 2, \dots, 10$, $i = 1, 2, \dots, 8$. The linear model is

$$x_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

The null hypothesis is $H_0 : \tau_1 = \tau_2 = \dots = \tau_8$.

(b) The alternative hypothesis is

$$H_1 : \tau_l \neq \tau_m \text{ for some } l \text{ and } m.$$

$$(c) N = 80, \bar{x} = \frac{59.01 + 59.22 + \dots + 60.23}{8} = 5.9455,$$

$$CF = 80 \times (5.9455)^2 = 2827.918, \sum \sum x_{ij}^2 = 2828.096,$$

$$TSS = 2828.096 - 2827.918 = 0.178,$$

$$SS_a = \frac{(59.01)^2 + \dots + (60.23)^2}{8} - 2827.918 = 0.099.$$

Table-13: ANOVA Slot Length Data

SV	DF	SS	MS	F
Stations	7	0.099	0.014	14
Error	72	0.079	0.001	
Total	79	0.178		

The tabulated F-value with 7 and 72 df is 2.13. Therefore, H_0 is rejected and we conclude that station means are not equal.

E7) Answers to (a) and (b) are same as in E6)

$$(c) N = 80, \bar{x} = \frac{59.1 + 59.03 + \dots + 59.11}{8} = 5.9191,$$

$$CF = 80 \times (5.9191)^2 = 2802.883, \sum \sum x_{ij}^2 = 2802.97,$$

$$TSS = 2802.97 - 2802.883 = 0.087, SS_a =$$

$$\frac{(59.1)^2 + \dots + (59.11)^2}{8} - 2802.883 = 0.011.$$

Table-14: ANOVA Slot Length Data

SV	DF	SS	MS	F
Stations	7	0.011	0.0015	1.5
Error	72	0.076	0.0010	
Total	79	0.087		

The tabulated F-value with 7 and 72 df is 2.13. Therefore, we cannot reject H_0 . There is no substantial evidence to conclude that station means are different.

E8) (a) $H_0 : \tau_1 = \tau_6$ Vs $H_1 : \tau_1 \neq \tau_6$.

$$LSD = t_{0.025, 72} \times \sqrt{2 \times MS_e / 10} = 1.99 \times \sqrt{2 \times 0.001 / 10} = 0.028.$$

Since the difference between \bar{x}_1 and \bar{x}_6 ($= 0.290$) is more than LSD, there is significant difference between the effects of station 1 and station 6.

(b) $H_0 : \tau_4 = \tau_6$ Vs $H_1 : \tau_4 \neq \tau_6$.

Since the difference between \bar{x}_4 and \bar{x}_6 ($= 0.200$) is more than LSD, we reject H_0 and conclude that the effects of station 4 and station 6 are significantly different.

(c) Assuming $\tau_4 = \tau_7 = \tau$, the estimate of $\mu + \tau$ is given by

$$u = \frac{\bar{x}_4 + \bar{x}_7}{2} = 5.9475.$$

The hypothesis to be tested is

$$H_0 : \mu + \tau = 5.95 \quad \text{Vs} \quad H_1 : \mu + \tau \neq 5.95.$$

Under normality assumptions, $\mu + \tau$ follows normal distribution with mean $\mu + \tau$ and variance $\sigma^2 / 20$. Under H_0 ,

$$t_0 = \frac{u - 5.95}{\sqrt{MS_e / 20}}$$

follows t-distribution with 72 df. Substituting the values, we get $t_0 = 0.353$. Since this is less than the tabulated t-value, we conclude that there is no substantial evidence to say that $\mu + \tau$ is away from the target.

(d) Let $\mu_3 = \mu + \tau_3$. We wish to test

$$H_0 : \mu_3 = 5.95 \quad \text{Vs} \quad H_1 : \mu_3 < 5.95.$$

Here we use the one-sided t-test and reject H_0 if

$$t_0 = \frac{\bar{x}_3 - 5.95}{\sqrt{MS_e / 10}}$$

is less than the lower 5% point of t-distribution with 72 df. Substituting the values,

$$t_0 = \frac{5.923 - 5.95}{\sqrt{0.001 / 10}} = -2.7.$$

Since the tabulated t-value is -1.66, we reject H_0 and conclude that μ_3 is significantly lower than the target.

- E9) (a) When all τ_i s are equal, say equal to τ , \bar{x} is an estimate of $\mu + \tau$ and has normal distribution with mean $\mu + \tau$ and variance $\frac{\sigma^2}{80}$. To test

$$H_0 : \mu + \tau = 5.95 \quad \text{Vs} \quad H_1 : \mu + \tau \neq 5.95$$

we compute the t-statistic

$$t_0 = \frac{\bar{x} - 5.95}{\sqrt{MS_e/80}} = \frac{5.919 - 5.95}{0.0035} = -8.76.$$

Conclude that the special purpose machine setting is significantly lower than the target value.

- E10) To test

$$H_0 : \tau_2 = \tau_3 \quad \text{Vs} \quad H_1 : \tau_2 \neq \tau_3$$

compute

$$t_0 = \frac{21.5 - 21.72}{\sqrt{0.3599/(\frac{1}{4} + \frac{1}{5})}} = -0.275.$$

Since the tabulated t-value is 2.145, there is no substantial evidence to reject H_0 .

- E11)

Table-15 : Summary of Tailors' Productivity Data

Summary	Tailor			Total
	A	B	C	
n_i	4	5	3	12
Total	25	35	23	83
Average	6.25	7	7.66	21.41
$\sum y_{ij}^2$	151	251	161	593

$$n = 12, CF = 574.0833.$$

Table-16: ANOVA For Tailors' Data

SV	DF	SS	MS	F
Tailors	2	3.50	1.75	1.5
Error	9	15.4166	1.7129	
Total	11	18.9166		

Tabulated F-value with 2 and 9 df at 5% level is equal to 4.256. There is no substantial evidence to say that the tailors are not equally productive.

Table-17: ANOVA For Fabric Strength Data

SV	DF	SS	MS	F
Tailors	2	81.5	40.75	29.34
Error	9	12.5	1.3888	
Total	11	94.0		

The estimates of variance components are given by

$$\hat{\sigma}^2 = 1.3888 \text{ and } \sigma_\tau^2 = \frac{40.75 - 1.3888}{4} = 9.8403.$$

UNIT 9 REGRESSION ANALYSIS

Structure	Page No.
9.1 Introduction Objectives	23
9.2 Simple Linear Regression	24
9.3 Measures of Goodness of Fit	31
9.4 Multiple Linear Regression Preliminaries Regression With Two Independent Variables	32
9.5 Summary	36
9.6 Solutions/Answers	36

9.1 INTRODUCTION

Decision making is an important activity of everybody's life. It requires knowledge and information. For example, for a farmer to decide whether to grow paddy or not in a particular field, he should know the soil conditions, availability of water, etc. An experienced farmer will be able to predict the yield of paddy by examining the soil features. In this example, the yield of paddy, y , is *associated* with the fertility of the soil, x . It would be nice if there is a reasonably good mathematical relationship that can be used to predict the yield, y , at least approximately, for a given value of x , so that the farmer can work out the yield and decide whether it is worth growing paddy or not. *Regression Analysis* is a powerful statistical technique useful in building such mathematical relationships.

To make you appreciate the subject, here are some interesting practical applications of the technique. These are compiled from project studies carried out in Indian industries.

An Application in a Sugar Factory

The yield of sugar depends on the time at which the sugar cane is harvested. During the harvesting period, the yield of sugar that can be obtained from the sugar cane gradually increases with time up to a certain period and starts falling beyond that period. To get maximum yield, the sugar factory has to decide when to harvest the crops. This decision making requires the relationship between the amount of yield and the time. Using past data, a regression equation was developed which formed the basis for arriving at a decision making procedure. This project was carried out in a sugar factory situated in Andhra Pradesh. As a result of this project, the factory's profits went up by several lakhs of rupees by way of increased turnover.

An Application in an Electronic Industry

In one semiconductor manufacturing unit, about 41% of the transistors produced were getting rejected due to various faults. Rejection data analysis revealed that 39% of the 41% rejections were due to low β -value while the remaining 2% were caused by other faults. The β -value, which should be between 30 and 100 according to specifications, was influenced by three process parameters x_1, x_2, x_3 (these are resistances at certain positions). It was possible to control these process parameters at desired levels. From the past data, the following formula was obtained using regression analysis.

$$\beta - \text{value} = 32.67 + 0.64x_1 - 1.18x_2 - 20.98x_3.$$

Using this equation, the levels of x_1, x_2, x_3 , which result in increased β -values, were identified and maintained in the process. As a result of this study, the rejections in transistors came down from 41% to 20%.

Prediction of Tensile Strength of Castings

For a particular grade of castings produced by a foundry, customers' specifications were in terms of hardness and tensile strength. While the hardness and chemical composition were being tested in the foundry, tensile strength was being tested by an outside laboratory. The results used to come after a time lag of 45 days and as such were useless from the view point of process control. Though it was known that tensile strength depends on the chemical composition and hardness, no relationship between these variables was available. Using the past data, the following regression equation was developed.

$$y = 19.366 - 8.359x_1 + 2.854x_2 + 10.675x_3 + 0.096x_4,$$

where y is tensile strength, x_1, x_2, x_3 are carbon, silicon and manganese percentages respectively, and x_4 is hardness. This prediction formula was discussed with foundry management who decided to use it for day to day process control.

The term regression was first used as a statistical concept by an English Statistician Sir Francis Galton in 1877. Galton made a study that showed that the heights of the children born to unusually tall or short parents tends to move back or "regress" towards the mean height of the population

Thus, you have seen three applications of regression analysis technique. It is one of the most widely applied statistical techniques. In this unit, you will learn this technique through illustrative examples. Through a simple example, we will learn how to build a simple *linear regression equation* to predict a given variable y based on another associated variable x . We will look at some simple methods to examine how good such an equation is. After this, we will consider the situations where we have two associated variables and learn how to build the formula and assess it. For example, you can think of predicting yield of paddy y , based on two associated variables: (i) x , the soil fertility and (ii) u , the quantity of chemical fertiliser applied. Finally, we will consider the case where three associated variables are involved.

Objectives

After reading this unit, you should be able to

- identify linear relationship between two variables through scatter diagram (a graphical aid),
- specify the simple linear regression model and build it from data,
- compute the coefficient of correlation and evaluate the regression formula through this coefficient,
- specify multiple linear regression models with two independent variables and build them from data,

9.2 SIMPLE LINEAR REGRESSION MODEL

In this section, we will study the simple linear regression model. Many of the concepts that you learn here will be useful when we deal with multiple regression models in Section 9.4.

Let us start with an example.

When the floor of a house is plastered with cement, you know that one should not walk on it until it gets set. The setting time is an important characteristic of cement and it is mandatory that any cement manufacturing company should adhere to certain specification limits. Once the cement is produced, its setting time cannot be changed. So the manufacturers should be in a position to predict the setting time well before it is produced. Fortunately, it is possible to do this because the setting time depends, to a large extent, on the chemical composition of the raw materials used to produce cement.

In Table-1, you will find data on 25 samples of cement. For each sample, we have a pair of observations (x, y) , where x is percentage of SO_3 , a chemical, and y is the setting time in minutes. Our aim is to study how y depends on x . We will refer to y as the **dependent variable or response**, and x as **independent variable or regressor**. You know that it is often easy to understand data through a graph. So, let us plot the data on a *scatter diagram*. A scatter diagram is a simple two-dimensional graph in which the horizontal axis represents x and the vertical axis represents y . Each pair of points is plotted on the graph. See the graph in Figure 1 given in the next page.

Table-1: Data on SO_3 and Setting Time

S. No. i	Percentage of SO_3 x	Setting Time y (in minutes)
1	1.84	190
2	1.91	192
3	1.90	210
4	1.66	194
5	1.48	170
6	1.26	160
7	1.21	143
8	1.32	164
9	2.11	200
10	0.94	136
11	2.25	206
12	0.96	138
13	1.71	185
14	2.35	210
15	1.65	178
16	1.19	170
17	1.56	160
18	1.53	160
19	0.96	140
20	1.67	168
21	1.68	152
22	1.28	160
23	1.35	116
24	1.49	145
25	1.78	170
Total	39.04	4217
Sum of Squares	64.446	726539

From this figure, you see that y increases as SO_3 increases. Whenever you find this type of increasing (or decreasing) trend in a scatter diagram, it indicates that there is a linear relationship between x and y . You may observe that the relationship is not perfect in the sense that a straight line cannot be drawn through all the points in the scatter diagram. Nevertheless, we may approximate it with some linear equation. What formula shall we use? Suppose we use the formula $y = 90 + 50x$ to predict y based on x . To examine how good this formula is, we need to compare the actual values of y with the corresponding predicted values. When $x = 0.96$, the predicted y is equal to $138 (= 90 + 50 \times 0.96)$. Let (x_i, y_i) denote the values of (x, y) for the i th sample. From Table-1, notice that $x_{12} = x_{19} = 0.96$ where as $y_{12} = 138$ and $y_{19} = 140$.

Let $\hat{y}_i = 90 + 50x_i$. That is, \hat{y}_i is the predicted value of y (using $y = 90 + 50x$) for the i th sample. Since, $x_{12} = x_{19} = 0.96$, both \hat{y}_{12} and \hat{y}_{19} are equal to 138. The difference $\hat{e}_i = y_i - \hat{y}_i$, the **error in prediction**, is called the **residual**. Observe that $\hat{e}_{12} = 0$ and $\hat{e}_{19} = 2$. The formula we have considered above, $y = 90 + 50x$, is called a **simple**

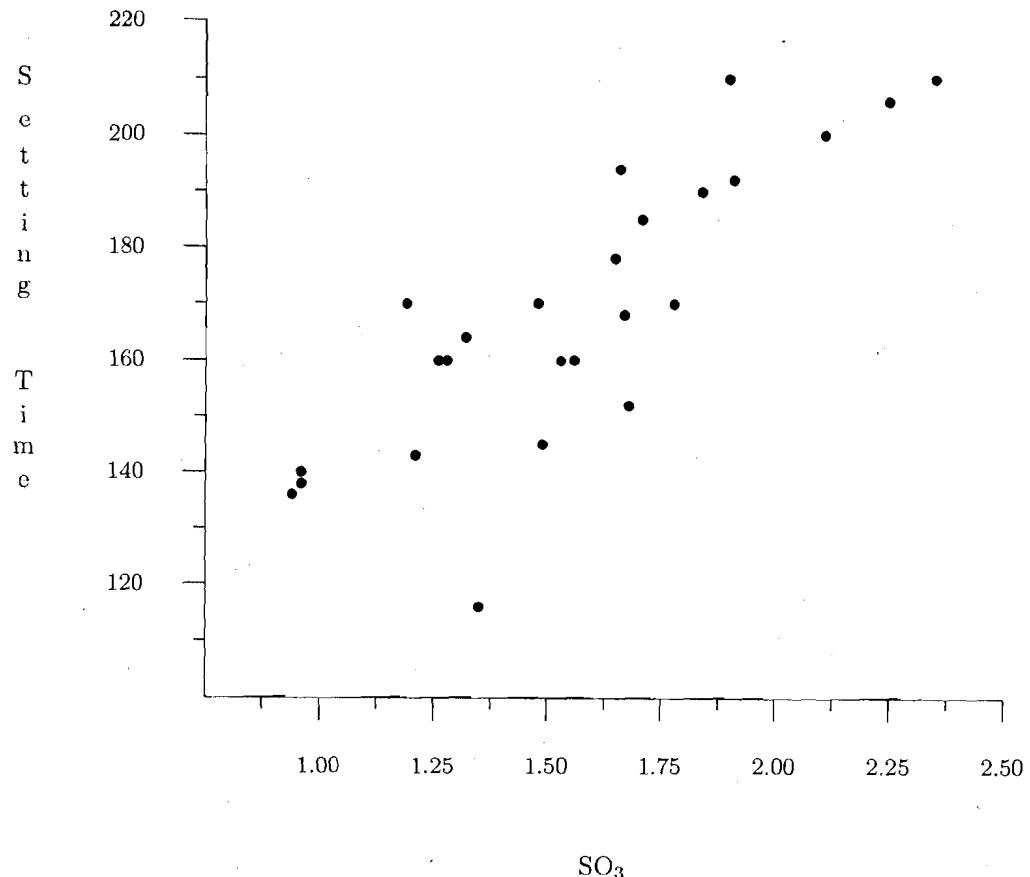


Fig. 1: Scatter Diagram of Setting Time vs SO_3

linear regression equation. If you recall from your study of coordinate geometry, this equation is the equation of a straight line. You can easily see that \hat{y}_i and \hat{e}_i depend on the straight line we choose to predict y . That is, if we change the equation, let us say, from $y = 90 + 50x$ to $y = 100 + 45x$, then the \hat{y}_i and \hat{e}_i will be different. To fix these ideas in your mind, try these exercises now.

E1) Using the regression equation $y = 90 + 50x$, fill up the values in the table below.

Table-2: \hat{y}_i and \hat{e}_i For Some Selected i

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
\hat{y}_i	138				
\hat{e}_i	0				

Note: $\hat{y}_i = 90 + 50x$ and $\hat{e}_i = y_i - \hat{y}_i$

E2) Using $y = 100 + 45x$, fill up the values in the table below.

Table-3: \hat{y}_i and \hat{e}_i For Some Selected i

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
\hat{y}_i	143.2				
\hat{e}_i	-5.2				

Note: $\hat{y}_i = 100 + 40x$ and $\hat{e}_i = y_i - \hat{y}_i$

- E3) Compare the \hat{y}_i values of Table-2 with the corresponding ones in Table-3. Do this comparison with respect to \hat{e}_i s also.

From the exercises above, you might have noticed that different equations give us different residuals. What will be the best equation? Obviously, the choice will be that equation for which \hat{e}_i s are small.

From samples 12 and 19, we have found that for the same value of $x = 0.96$, the y values are different. This means that whatever straight line we use, it is not possible to make all the \hat{e}_i s zero. However, we would expect that the errors are positive in some cases and negative in the other cases so that, on the whole, their sum is close to zero. So, our job is to find out the *best* values of a and b in the formula $y = a + bx$. Let us see how we do this.

Now our aim is to find the values a and b so that the error e_i s are minimum. For that we state here four steps to be done.

- 1) Calculate a sum S_{xx} defined by

$$S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \quad (1)$$

where x_i 's are the given values of the data and $\bar{x} = \frac{\sum x_i}{n}$ is the mean of the observed values and n is the sample size.

The sum S_{xx} is called the corrected sum of squares.

- 2) Calculate a sum S_{xy} defined by

$$S_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \quad (2)$$

where x_i 's and y_i 's are the x -values and y -values given by the data and \bar{x} and \bar{y} are their means.

- 3) Calculate $\frac{S_{xy}}{S_{xx}} = b$, say. That is

$$b = \frac{S_{xy}}{S_{xx}} \quad (3)$$

- 4) Find $\bar{y} - b\bar{x} = a$, say.

Let us now compute these values for the data in Table-1, we get

$$\bar{x} = 1.5616, \bar{y} = 168.68, S_{xx} = 3.4811, \text{ and } S_{xy} = 191.2328.$$

Substituting these values in (3) and (4), we get

$$b = \frac{S_{xy}}{S_{xx}} = 54.943 \text{ and } a = 168.68 - 54.943 \times 1.5616 = 82.88. \quad (4)$$

Therefore, the best linear prediction formula is given by

$$y = 82.88 + 54.943x. \quad (5)$$

So far, we have come across three prediction formulae, namely, $y = 90 + 50x$, $y = 100 + 45x$ and $y = 82.88 + 54.943x$. While the first two of these were proposed in an ad hoc manner, the third one was obtained objectively *using data*. Let us draw these three lines on the scatter diagram and see how the picture looks like. From Figure 2 on page number 28, you can see that compared to straight lines in (a) and (b), the one in (c) is close to more points.

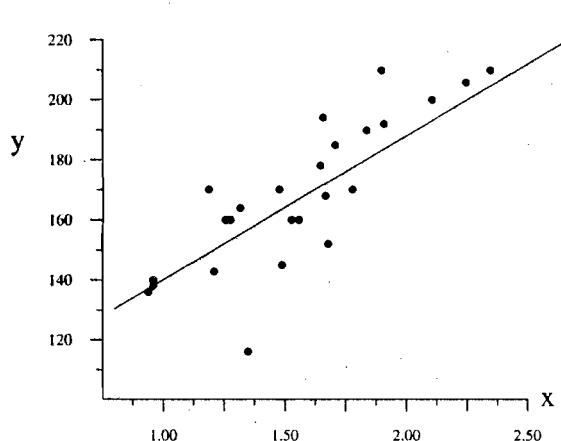
In the exercise below, we have another set of observations on (x, y) . Now do this exercise to make sure that you have understood the procedure to get best linear regression formula.

- E4) Work out the best linear regression formula from the following data.

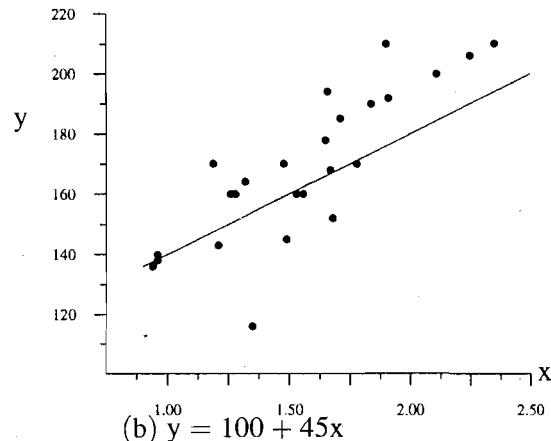
Table-4: Data on SO₃ and Setting Time

S. No. (i)	Percentage SO ₃ (x)	Setting Time y (in minutes)	S. No. (i)	% SO ₃ (x)	Setting Time y (in minutes)
1	2.25	206	9	1.67	168
2	0.96	138	10	1.28	160
3	1.71	185	11	1.78	170
4	2.35	210	12	0.78	130
5	1.65	178	13	1.26	146
6	1.56	160	14	1.50	180
7	1.53	160	15	1.35	148
8	0.96	140			
$\sum x_i = 22.59, \sum x_i^2 = 36.7135, \sum y_i = 2479, \sum y_i^2 = 511637.$					

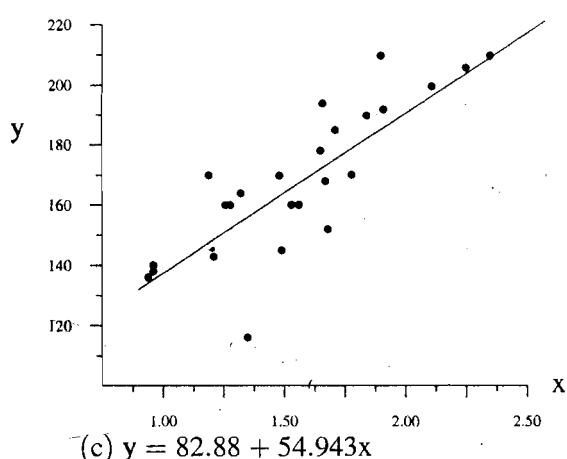
If you have done the above exercise, you must have got $y = 88.05 + 51.273 x$ as your best linear regression formula. Compare this with what we have got earlier (see Eqn. (5)). Have you noticed that the values of a and b have changed. Well, this is bound to happen because the set of observations we have used there was different from the one



(a) $y = 90 + 50x$



(b) $y = 100 + 45x$



(c) $y = 82.88 + 54.943x$

we have used in (E4). The point that you should learn from this exercise is that the values of a and b depend on the sample of observations. Therefore, it would have been proper for us to have used the symbols \hat{a} and \hat{b} in Eqns.(4) and (6) to indicate that these are the estimated values obtained from the sample of observations.

By now you must have understood how to calculate the best linear regression line and how will we use it to make some predictions about the problems that are analysed. Let us see an example.

Problem 1: A hosiery mill wants to estimate how its monthly costs are related to its monthly output rate. For that the firm collects a data regarding its costs and output for a sample of nine months as given in Table 5 below.

Table 5

Output (tons)	Production cost (thousands of dollars)
1	2
2	3
4	4
8	7
6	6
5	5
8	8
9	8
7	6

- 1) Construct a scatter diagram for the data given above.
- 2) Calculate the best linear regression line, where the monthly output is the dependent variable and the monthly cost is the independent variable.
- 3) Use this regression line to predict the firm's monthly costs if they decide to produce 4 tons per month.

Solution:

- 1) Suppose that x_i denote the output for the i th month and y_i denote the cost for the i th month. Then we can plot the graph for the pair (x_i, y_i) of the values given in Table 5. Then we get the scatter diagram as shown in Fig. 3 in the next page.
- 2) Now to find the least square regression line, we first calculate the sums S_{xx} and S_{xy} from Eqn.(1) and (2).

$$S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}$$

Note that from Table (5) we get that

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{50}{9}$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{49}{9}$$

$$\sum x_i^2 = 340$$

$$\sum y_i^2 = 303$$

$$\text{and } \sum x_i y_i = 319$$

Therefore we get that

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

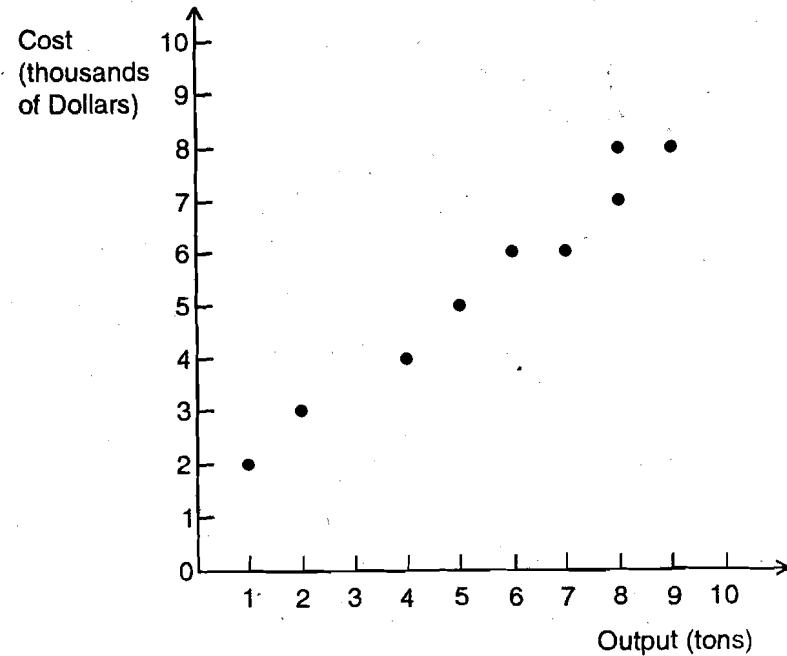


Fig. 3: Scatter Diagram

$$\begin{aligned}
 &= \frac{9 \times 319 - 50 \times 49}{9 \times 340 - 50^2} \\
 &= \frac{421}{560} = 0.752
 \end{aligned}$$

Correspondingly, we get

$$\begin{aligned}
 \hat{a} &= \frac{49}{9} - (0.752) \times \frac{50}{9} \\
 &= 1.266
 \end{aligned}$$

Therefore the best linear regression line is

$$y = 1.266 + (0.752)x$$

- 3) If the firm decides to produce 4 tons per month, then one can predict that its cost would be

$$1.266 + (0.752) \times 4 = 4.274$$

Since the costs are measured in thousands of dollars, this means that the total costs would be expected to be \$4,274.

————— X —————

The above example illustrates that the regression line can be of great practical importance. You will be more clear about this, when you try the following exercise.

-
- E5) An economist wants to estimate the relationship in a small community between a family's annual income and the amount that the family saves. The following data from nine families are obtained:

Annual income (thousands of dollars)	Annual savings (thousands of dollars)
12	0.0
13	0.1
14	0.2
15	0.2
16	0.5
17	0.5
18	0.6
19	0.7
20	0.8

Calculate the least-squares regression line, where annual savings is the dependent variable and annual income is the independent variable, and interpret your results.

In the next section we shall consider some procedures to measure how good our best fit is.

9.3 MEASURES OF GOODNESS OF FIT

You have seen that regression line provides estimates of the dependent variable for a given value of the independent variable. The regression line is called the line of best fit. It shows the relationship between x and y better than any other line.

Let us consider the question "How good is our best linear regression formula?" We would like to able to measure how good our best fit is. More precisely we want to have some measures for goodness of fit.

To develop a measure of goodness of fit, we first examine the variation in y . Let us try to examine the variation in the response y . Since y depends on x , if we change x , then y also changes. In other words, a part of variation in y_i s is accounted by the variation in x_i s. Actually, we can mathematically show that the total variation in y_i s can be split up as follows:

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{S_{xy}^2}{S_{xx}} + \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (6)$$

Dividing this equation by S_{yy} on both sides, we get

$$1 = \frac{S_{xy}^2}{S_{xx} S_{yy}} + \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{S_{yy}}. \quad (7)$$

Since the quantities on the right hand side are both nonnegative, none of them can exceed one. Also, if one of the two is closer to one, then the other has to be closer to zero. Let us use the notation

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}. \quad (8)$$

Then, from what we have just now argued, $R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$ must be between 0 and 1.

This, in turn, will imply that R will be between -1 and +1.

Supposing $R^2 = 1$, then from (7) we see that $\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{S_{yy}} = 0$ or $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 0$

or $y_i = \hat{y}_i$ for all i . Again, when R^2 is close to 1, $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is close to zero. R is

called the correlation coefficient between x and y . When R is negative, it means that y decreases as x increases; and when R is positive, y increases as x increases. Thus, R gives a measure of the strength of the relationship between the variables x and y .

Let us compute R and R^2 for the data in Table-1.

$$R = \frac{191.2328}{\sqrt{3.4807 \times 15216.7776}} = 0.8309 \text{ and } R^2 = 0.6904.$$

By now, you must have understood that the large R^2 (i.e., closer to 1) should mean that our regression formula is more reliable in the sense that predicted values from such a regression equation will be close to the actual values.

Let us now look at the value of the correlation coefficient R obtained for the data in Table 1. Since, in this case, $R^2 = 0.69$ is not zero, we can conclude on the basis of our sample data that there is a relationship between x and y.

Now the question is: How large should R^2 be. We can answer this question by carrying out a statistical test. But before we go onto that, try this exercise.

E6) For the data in Table 2, compute R and R^2 . Compare this with the above result.

In order to carry out statistical tests, we shall assume that e_i s are independent and normally distributed. Coming to the question of what values of R^2 can be considered as large, it is logical to compare R^2 with $1 - R^2$ (in the light of the fact that R^2 can be at most equal to 1). In fact, it has been shown that $\frac{(n - 2)R^2}{1 - R^2}$ follows F-distribution with 1 and $(n - 2)$ degrees of freedom (df). For the data in Table-1, this value is given by

$$\frac{(n - 2)R^2}{1 - R^2} = \frac{23 \times 0.6904}{1 - 0.6904} = 51.29.$$

From the table of F-distribution given in the Appendix, we get that the tabulated F-value at 5% level of significance (i.e. $\alpha = 0.05$) with 1 and 23 df is equal to 4.28. Therefore in this example, R^2 is significantly large.

So far in this unit we have considered only one independent variable and a dependent variables. Many situations require the use of more than one independent variable to explain the values that the dependent variable may take. In the next section we shall discuss this.

9.4 MULTIPLE LINEAR REGRESSION

In the previous section we considered problems where the response y depended on a single independent variable x. But, often responses depend on several independent variables. For example, the yield of paddy not only depends on soil fertility but also on the amount of water used, the quantity of chemical fertilisers applied and so on. In this section we will consider problems in which response depends on 2 or 3 independent variables. Many of the ideas and concepts that we have learnt in simple linear regression model are extended to multiple linear regression as well. The only change is that the computations will be a little more involved. Here we shall explain how the techniques are extended in the multiple case.

First we shall give some preliminaries for formulating the regression line.

9.4.1 Preliminaries

We shall learn the multiple regression technique with the help of a study that was carried out in a spring manufacturing company in Bangalore a few years ago.

Example 1: One of the types of springs manufactured by the company is known as flat type springs which are used in textile machines. These springs are produced in batches and after production a spring is selected at random from each batch and is subjected to vibration test. The spring has to withstand 1800 vibrations, failing which the batch will

be sent for rework. At one time the company was facing severe rework problem. Most of the batches were getting rejected in the vibration test, and hence, were being reworked.

Tempering is one of the operations in the manufacture of springs in which springs are heated to a particular temperature, known as tempering temperature, and are kept at that temperature for a specified amount of time, known as tempering time. After tempering the springs are soaked in a chemical solution for a specified amount of time which is known as soaking time. We shall denote the tempering temperature by x_1 , tempering time by x_2 and soaking time by x_3 .

After preliminary investigations it was found out that there was lack of control in x_1 , x_2 and x_3 . It was not clear whether the variation in these variables was causing failures in the vibration test. To study the problem, data were collected for 15 batches, and for each batch the values of x_1 , x_2 , x_3 and y , the number of vibrations (in hundreds) withstood by the spring selected at random from the batch. These data are presented in Table-6.

Table-6: Flat Springs Data

Sl. No.	Tempering		Soaking Time (x_3)	Number of Vibrations (in 100s) (y)
	Temperature (x_1)	Time (x_2)		
1	321	59	26	19.01
2	339	69	24	12.66
3	334	63	27	17.45
4	329	70	20	12.32
5	325	58	23	19.15
6	331	56	24	19.09
7	324	61	27	17.98
8	321	69	29	18.50
9	337	59	22	17.48
10	335	68	20	12.93
11	320	69	28	16.80
12	331	64	22	11.79
13	329	67	30	20.53
14	336	60	21	14.32
15	339	58	25	19.21

We shall now analyse the data and see how we will resolve the problem of high rework. In this process, we shall learn the multiple linear regression technique with two and three independent variables. To ease your learning process, it is better to compute certain quantities before going into multiple regression. Let us know what these quantities are and how to compute them. In Table-6 we have 15 observations on (x_1, x_2, x_3, y) . The i^{th} observation on x_1 will be denoted by x_{1i} . Similar notation will be used for x_2 , x_3 and y . Each of the 15 observations on (x_1, x_2, x_3, y) is called a data point. Thus, $(325, 58, 23, 19.15)$ is the fifth data point.

The **corrected sum of products** between the variables x_1 and x_2 is denoted by S_{12} and is defined by

$$S_{12} = \sum_{i=1}^{15} x_{1i}x_{2i} - \frac{(\sum_{i=1}^{15} x_{1i})(\sum_{i=2}^{15} x_{2i})}{15} \quad (9)$$

The **corrected sum of products** between the variables x_i and y is denoted by S_{iy} and is defined as

$$S_{iy} = \sum_{i=1}^{15} x_{1i}y_i - \frac{(\sum_{i=1}^{15} x_{1i})(\sum_{i=2}^{15} y_i)}{15}$$

The corrected sum of squares of x_1 is denoted by S_{11} and is defined by

$$S_{11} = \sum_{i=1}^{15} x_{1i}^2 - \frac{(\sum_{i=1}^{15} x_{1i})^2}{15}$$

* * *

Note that when only one variable is involved, we call it sum of squares, and if two variables are involved, we call it sum of products. It will be easy for you to do the below exercise.

E7) Write down the notation for the following and define them for the data given in Table 6.

- (i) corrected sum of products between x_2 and x_3
- (ii) corrected sum of products between x_3 and y
- (iii) corrected sum of squares of y .

E8) Explain the following: (i) S_{13} , (ii) S_{33} .

Example 2: Let us now compute the quantities, S_{11} , S_{12} , S_{22} , S_{1y} and S_{2y} for the data given in Table 6.

Forming Table-7 and Table-8 will make our computations easier.

**Table-7 : Computation of Sum
of Squares and Products**

S.No.	x_1	x_2	x_1^2	x_2^2	x_1x_2
1	321	59	103041	3481	18939
2	339	69	114921	4761	23391
3	334	63	111556	3969	21042
4	329	70	108241	4900	23030
5	325	58	105625	3364	18850
6	331	56	109561	3136	18536
7	324	61	104976	3721	19764
8	321	69	103041	4761	22149
9	337	59	113569	3481	19883
10	335	68	112225	4624	22780
11	320	69	102400	4761	22080
12	331	64	109561	4096	21184
13	329	67	108241	4489	22043
14	336	60	112896	3600	20160
15	339	58	114921	3364	19662
Total	4951	950	6134775	60508	313493

**Table-8 : Computation of Sum
of Squares and Products**

S.No.	x_1	x_2	y	y^2	x_1y	x_2y
1	321	59	19.01	361.3801	6102.21	1121.59
2	339	69	12.66	160.2756	4291.74	873.54
3	334	63	17.45	304.5025	5828.30	1099.35
4	329	70	12.32	151.7824	4053.28	862.40
5	325	58	19.15	366.7225	6223.75	1110.70
6	331	56	19.09	364.4281	6318.79	1069.04
7	324	61	17.98	323.2804	5825.52	1096.78
8	321	69	18.50	342.2500	5938.50	1276.50
9	337	59	17.48	305.5504	5890.76	1031.32
10	335	68	12.93	167.1849	4331.55	879.24
11	320	69	16.80	282.2400	5376.00	1159.20
12	331	64	11.79	139.0041	3902.49	754.56
13	329	67	20.53	421.4809	6754.37	1375.51
14	336	60	14.32	205.0624	4811.52	859.20
15	339	58	19.21	369.0241	6512.19	1114.18
Total	4951	950	249.22	4264.168	82160.97	15683.11

From Table-7,

$$S_{11} = 1634775 - \frac{(4951)(4951)}{15} = 614.9333,$$

$$S_{12} = 313493 - \frac{(4951)(950)}{15} = -70.3333,$$

$$S_{1y} = 82160.97 - \frac{(4951)(249.22)}{15} = -98.2446.$$

* * *

You can now easily do the following exercises.

E9) Compute S_{yy} and S_{2y} .

E10) Guess what S_{21} will be. Is it same as S_{12} ? What can you say about S_{1y} and S_{y1} ?

E11) Compute S_{13} , S_{23} , S_{33} and S_{3y} . To save your time the following quantities are already computed for you.

$$\sum x_{1i}x_{3i} = 121313, \sum x_{2i}x_{3i} = 23335, \sum x_{2i}y_i = 15683.11,$$

$$\sum x_{3i} = 368, \sum x_{3i}y_i = 6206.03, \sum x_{3i}^2 = 9174.$$

You are now in a position to learn the multiple regression technique easily. Here, we shall look at multiple regression with two independent variables only.

9.4.2 Regression With Two Independent Variables

Recall the spring rework problem. For the time being let us ignore the soaking time and examine the effect of tempering temperature (x_1) and tempering time (x_2) on vibrations, y . This is done by fitting a linear regression formula for y on x_1 and x_2 . The form of the linear regression function is given by

$$y = a_0 + a_1x_1 + a_2x_2 + e,$$

where a_0 , a_1 and a_2 are unknown constants and e is the random error with mean zero and standard deviation σ . We have 15 observations on this model, namely, the 15 data points (x_{1i}, x_{2i}, y_i) presented in Table 8 (note that x_{3i} s are omitted as we are not considering x_3). As we have done in the one independent variable case, a_0 , a_1 and a_2 are estimated by minimising the error sum of squares

$$SSE = \sum_{i=1}^{15} e_i^2 = \sum_{i=1}^{15} (y_i - a_0 - a_1x_{1i} - a_2x_{2i})^2.$$

The constants a_0 , a_1 , a_2 and σ^2 are estimated from the data. Let these estimates be denoted by \hat{a}_0 , \hat{a}_1 , \hat{a}_2 and $\hat{\sigma}^2$. The resulting estimates \hat{a}_1 and \hat{a}_2 are obtained by solving the following equations:

$$S_{11} a_1 + S_{12} a_2 = S_{1y}$$

$$S_{21} a_1 + S_{22} a_2 = S_{2y}$$

Above, we have already computed S_{11} , S_{12} etc., from the data. Substituting these values, we get

$$614.9333 a_1 - 70.3333 a_2 = -98.2447$$

$$-70.3333 a_1 + 341.3333 a_2 = -100.8230$$

We can solve these equations for a_1 and a_2 to obtain \hat{a}_1 and \hat{a}_2 respectively. However, there is a simple formula for \hat{a}_1 and \hat{a}_2 when $\Delta = S_{11}S_{22} - S_{12}^2 \neq 0$ which is the case in all most all the examples that we come across in practice. In our example,

$$\Delta = 614.9333 \times 341.3333 - (-70.3333)^2 = 204950.44 \neq 0$$

The estimates of \hat{a}_1 and \hat{a}_2 are given by

$$\hat{a}_1 = C_{11}S_{1y} + C_{12}S_{2y} \quad (10)$$

$$\hat{a}_2 = C_{21}S_{1y} + C_{22}S_{2y} \quad (11)$$

where $C_{11} = S_{22}/\Delta$, $C_{12} = C_{21} = -S_{12}/\Delta$ and $C_{22} = S_{11}/\Delta$. Let us compute these quantities now.

$$C_{11} = \frac{341.3333}{204950.44} = 0.0017, C_{22} = \frac{614.9333}{204950.44} = 0.0030,$$

$$C_{12} = C_{21} = -\frac{-70.3333}{204950.44} = 0.0003.$$

Substituting C_{ij} s and S_{iy} s in Equations (10) and (11) we get

$$\hat{a}_1 = -0.1982$$

$$\hat{a}_2 = -0.3362,$$

To get the regression equation we need to obtain \hat{a}_0 . This is given by the formula

$$\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}_1 - \hat{a}_2\bar{x}_2, \quad (12)$$

where \bar{y} , \bar{x}_1 and \bar{x}_2 are the averages of y , x_1 and x_2 computed from the data. Thus,

$$\hat{a}_0 = 16.615 - (-0.1982) \times 330.07 - (-0.3362) \times 63.33 = 103.33.$$

Therefore, our linear regression formula for vibrations y based on x_1 and x_2 is given by

$$y = 103.33 - 0.1982 x_1 - 0.3362 x_2. \quad (13)$$

From the regression equation, we find that the coefficients of both x_1 and x_2 are negative. This means that number of vibrations can be increased by reducing the tempering temperature and tempering time or both.

Try this exercise now.

E12) Suppose we maintain x_1 at 320°C and x_2 at 60 minutes, then find the expected number of vibrations that a spring will withstand.

With this, we bring this unit to a close. Let us go back and recall the points covered in it.

9.5 SUMMARY

In this unit you have seen

- 1 that, regression analysis is a very useful technique by looking at some case applications,
- 2 how to identify linear relationship between a dependent variable and an independent variable by examining the scatter diagram,
- 3 the simple linear regression formula and how to build it from the data using the method of least squares principle,
- 4 the correlation coefficient and its significance in evaluating the regression formula,
- 5 the multiple linear regression formula with two independent variables and how to build them from data.

9.6 SOLUTIONS/ANSWERS

E1)

Table-2: \hat{y}_i and \hat{e}_i For Some Selected i

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
\hat{y}_i	138	154	172.5	182	207.5
\hat{e}_i	0	6	5.5	8	2.5

Note: $\hat{y}_i = 90 + 50x$ and $\hat{e}_i = y_i - \hat{y}_i$

E2)

Table-3: \hat{y}_i and \hat{e}_i For Some Selected i

Sample No (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
\hat{y}_i	143.2	157.6	174.25	182.8	205.75
\hat{e}_i	-5.2	2.4	3.75	7.2	4.25

Note: $\hat{y}_i = 100 + 40x$ and $\hat{e}_i = y_i - \hat{y}_i$

E3) Most of the errors are on the positive side with both the formulae.

$$\begin{aligned} E4) \quad S_{xx} &= 36.7135 - \frac{(22.59)^2}{15} = 2.6929, \sum x_i y_i = 3871.45, S_{xy} = 138.0760, \\ S_{yy} &= 417533 - \frac{(2479)^2}{15} = 7836.93. \\ \hat{b} &= \frac{S_{xy}}{S_{xx}} = \frac{138.076}{2.6929} = 51.2729, \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} = 165.2666 - 51.2729 \times 1.506 = 88.0496. \end{aligned}$$

Therefore, the best linear regression formula is

$$y = 88.0496 - 51.2729 x.$$

E5) Letting x_i be the income (in thousands of dollars) of the i th family, and y_i be the saving (in thousands of dollars) of the i th family, we find that

$$\begin{aligned} \sum_{i=1}^9 x_i y_i &= 63.7, \sum_{i=1}^9 y_i = 3.6 \quad \bar{v} = 0.4, \\ \sum_{i=1}^9 x_i^2 &= 2364, \quad \sum_{i=1}^9 x_i = 144, \quad \bar{x} = 16. \end{aligned}$$

Thus, substituting these values in the alternate formula for \hat{b} , we obtain

$$\hat{b} = \frac{9(63.7) - (144)(3.6)}{9(2364) - 144^2} = \frac{573.3 - 518.4}{21.276 - 20.736} = 0.1017.$$

Consequently,

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 0.4 - 0.1017(16) = -1.2272.$$

Thus, the regression line is

$$y = -1.2272 + 0.1017x,$$

where both x and y are measured in thousands of dollars.

The interpretation of this regression line is as follows: On the average, families with zero income would be expected to save -1,227.20. (Negative saving means that a family's consumption expenditure exceeds its income.) An increase in family income of 1,000 is associated with an increase in family saving of 101.70.

$$E6) \quad R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{138.076}{\sqrt{2.6929 \times 7839.93}} = 0.9504$$

$$\therefore R^2 = 0.9033$$

There is a substantial increase in R^2 .

E7) i) The notation for the corrected sum of products between x_2 and x_3 is S_{23} and this is defined by

$$S_{23} = \sum_{i=1}^{15} x_{2i} x_{3i} - \frac{(\sum_{i=1}^{15} x_{2i})(\sum_{i=1}^{15} x_{3i})}{15}.$$

ii) The notation for the corrected sum of products between x_3 and y is S_{3y} and this is defined by

$$S_{3y} = \sum_{i=1}^{15} x_{3i} y_i - \frac{(\sum_{i=1}^{15} x_{3i})(\sum_{i=1}^{15} y_i)}{15}.$$

iii) The notation for the corrected sum of squares of y is S_{yy} and this is defined by

$$S_{yy} = \sum_{i=1}^{15} y_i^2 - \frac{(\sum_{i=1}^{15} y_i)^2}{15}.$$

E8) (i) S_{13} is called the corrected sum of products between x_1 and x_3 ,

(ii) S_{33} is called the corrected sum of squares of x_3 .

E9) $S_{yy} = 4264.168 - \frac{(249.22)^2}{15} = 123.461,$

$$S_{2y} = 15683.11 - \frac{(950)(249.22)}{15} = -100.8233.$$

E10) S_{12} and S_{21} are equal as $x_{1i}x_{2i} = x_{2i}x_{1i}$. Similarly, $S_{1y} = S_{y1}$.

E11) $S_{13} = 151.533$, $S_{23} = 28.3333$, $S_{33} = 145.7333$ and $S_{3y} = 91.8326$.

E12) The required number is obtained by substituting the values $x_1 = 320^\circ\text{C}$ and $x_2 = 60 \text{ min.}$ in Eqn.(13). Then we get

$$103.33 - 0.1982 \times 320 - 0.3362 \times 62 = 19.62.$$

UNIT 10 FORECASTING AND TIME SERIES ANALYSIS

Structure	Page Nos.
10.1 Introduction	39
Objectives	40
10.2 Forecasting	40
10.3 Time Series and Their Components	43
Long-term Trend	45
Seasonal Variations	46
Cyclic Variations	48
Random Variations/Irregular Fluctuations	49
10.4 Forecasting Models	49
The Additive Model	50
The Multiplicative Model	52
10.5 Forecasting Long-term Trends	54
The Method of Least Squares	54
The Method of Moving Averages	58
Exponential Smoothing	62
10.6 Summary	69
10.7 Solutions and Answers	69

10.1 INTRODUCTION

In the previous unit you have seen how regression analysis was useful in decision making. If you carefully analyse most of our decisions and actions, be it by the Government, an institution, an industrial organisation or by an individual, they will greatly depend on the situation expected in future. For example, suppose the Government is thinking of a housing policy by which it will provide houses to all families of the fisher folk in Mumbai over the next five years. Then the Government must be able to assess what the number of families of fisher folk in Mumbai in the next five years would be. Similar assessments will be required in the case of designing an unemployment policy, and so on. An educational institution must assess the future needs so that it can start building up the required infrastructure, employ new teachers, and so on. A car manufacturer or a cement factory owner should be in a position to predict the future demands for her/his products so that she/he can plan the production schedules in the most profitable way. A person buying a real estate property will be benefitted by the knowledge on how the land value has been appreciating in the locality where she is planning to buy a piece of land.

Notice that in all the examples above, our interest lies in ‘predicting’ what the prevailing situation will be in future regarding various aspects. This kind of prediction is known as ‘forecasting’. This is largely based on the past behaviour of the particular aspect being studied.

In this unit you will study the methods of forecasting using time series. Data related to a particular feature, arranged in chronological order, is known as time series. You will learn some simple methods of time series data analysis and how to use the analysis in forecasting. In Sec.10.2 and Sec.10.3, we shall give formal definitions of forecasting and time series

and look at some examples of their applications. In Sec. 10.4, you will be introduced to basic components of a time series data and to the additive and multiplicative models. Our main focus in the analysis of time series data will be on understanding the long-term trend. Some methods of such trend analysis will be discussed in Sec. 10.5.

Before studying this unit, we expect you to have gone through Unit 9 thoroughly. Many concepts and formulae used there will be applied here too.

Objectives

After studying this unit, you will be able to

- define forecasting and justify its need;
- define time series data and its four basic components;
- explain the additive and multiplicative models;
- fit the linear (and some simple nonlinear) trends using the method of least squares;
- fit the trend by the method of moving averages;
- build forecasting models using simple exponential smoothing;
- briefly describe the Holt-Winter double parameter exponential smoothing model.

10.2 FORECASTING

All of us read or listen to the weather forecast often enough, wondering whether rain or a storm is going to hit us just when we don't want it to. The forecast, as you know, is a prediction of future conditions based on an analysis of data received over a period of time.

There are other kinds of forecasting too. For example, the Government may be interested in predicting the extent of damage due to natural calamities in the year ahead so that sufficient stock of foodgrains and other necessary commodities may be procured and reserved for such emergencies.

Forecasting can be used for predicting qualitative as well as quantitative aspects of events such as those described above. However, in this unit, we shall confine our discussion to the prediction of quantitative characteristics such as extent of damage quantified in rupees, the amount of food-grains required, etc. We call the aspect or condition we want to forecast, the **characteristic of interest**. These characteristics are generally influenced by a variety of factors such as economic conditions, technological advancements, population, inflation rates, weather conditions, seasonal effects, and so on. In most cases, it may not be possible to identify all these factors which affect the characteristic of interest. So we can't use them in predicting the future values of the characteristic. Instead, we apply a process of studying the behaviour of the characteristic over time for forecasting the expected values at certain points of time in the future. Typically, the forecasting methods have 2 parts — observing the pattern of the data on the characteristic over a time period; and then predicting the future expected values by assuming that the same pattern will continue in the future. So, the accuracy of such forecasts depends heavily on the

quality of the identification of the pattern and the validity of the assumption that the same pattern will continue in future. It is reasonable to assume in many cases that at least in the short run, the underlying pattern may not change. So this is a reasonable assumption for short-term forecasting. However, in the long run these assumptions can bring in errors. Such forecasting errors can be measured by computing the differences between actual values and the predicted values, and corrections be applied in future.

Let us examine how forecasting will be useful and how to compute the forecasting errors, through an example.

Example 1 : A grocery shop owner in a small colony gets 50 bread loaves every morning from a company and sells them to the residents of that colony. If he sells a loaf, he makes a profit of Rs.2/- . If a loaf is not sold on the same day, the shop owner returns it to the company the next morning but he loses Re.1/- on each loaf returned. In order to maximise his profits, the shop owner decided to study the pattern of the demand for the bread. (The daily demand for bread can be thought of as a random variable.) He collected the data shown in Table 1.

Table 1 : No. of Bread Loaves Demanded

Day	1st Week	2nd Week	3rd Week
Sunday	47	51	49
Monday	54	49	54
Tuesday	44	40	46
Wednesday	46	40	41
Thursday	63	59	57
Friday	51	56	50
Saturday	55	54	58

Let us now analyse this data. If we compute the average number of loaves that can be sold per day (that is, the demand per day) based on the three weeks' data, it comes to 50.66 loaves (= average of the 21 numbers in Table 1). In fact, from his past experience the shop owner found that he can sell approximately 50 loaves on an average. This is why he takes 50 loaves every morning from the company for selling.

So, based on the data gathered for 3 weeks, the owner's forecast for each day's sale is 50. Let us now compute the forecasting errors (for the owner's forecast) during the first 2 weeks (see Table 2).

Table 2 : Forecasting Errors For The First Two Weeks

Day	First Week			Second Week		
	Actual Demand	Predicted Demand	Forecasting Error	Actual Demand	Predicted Demand	Forecasting Error
Sun	47	50	-3	51	50	1
Mon	54	50	4	49	50	-1
Tue	44	50	-6	40	50	-10
Wed	46	50	-4	40	50	-10
Thurs	63	50	13	59	50	9
Fri	51	50	1	56	50	6
Sat	55	50	5	54	50	4

To check whether you have grasped the discussion so far, please try the following exercise now.

E1) Compute the forecasting errors for the third week.

Observe from Table 2, on Sunday of the first week the shop owner had to return 3 loaves, so he must have lost Rs.3/- . On Monday of the first week he fell short by 4 loaves. Had he had these loaves, he could have made an extra Rs.8/- profit. So this should also be considered as a loss resulting from his decision to get only 50 loaves every day. Therefore, the owner's profit on Sunday of the first week is equal to $(47 \times 2) - (3 \times 1) = 91$ rupees; his profit on Monday of the first week is equal to $(50 \times 2) - (4 \times 2) = 92$ rupees. In this way, if you work out the overall profit for the three weeks, it comes to Rs.1832/- (how?).

Do you think the owner can make a better decision so as to increase his profit? This may be possible if there is a way of making better forecasts of the sales. So, how can the owner improve upon his forecasting method so that the forecasting errors are smaller? To answer this, let us take another look at the data of Table 1. If we plot each week's data on a graph where the x-axis represents the days and the y-axis represents the demand, we get Fig. 1.

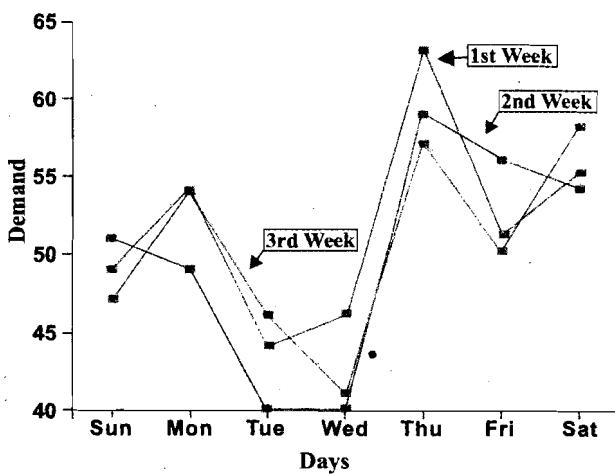


Fig. 1 : Demand of bread loaves vs. days.

From Fig. 1 we find the following trends :

- i) the demand varies from day to day;
- ii) on Tuesdays and Wednesdays the demand is lower than on other days of the week;
- iii) the demand seems to be significantly higher on Thursdays than on other days.

Remark : Note that the daily demand is a random variable which is dependent on time, and its expected values in future depend on the observed values in the past.

Don't you think it will be wiser on the part of the owner, in the light of the observations above, to make separate forecasts for each day of the week and then make his decision accordingly? How would he do this? One

way of doing so is to forecast each day's demand as the average of the three observations of that day. For example, the forecast for the demand on Sundays will be the average of 47, 51 and 49, which is equal to 49. Similarly, the forecast for the demand on Mondays will be the average of 54, 49 and 54, which is equal to 52.3. Since the number of loaves cannot be equal to 52.3, the owner may decide to get 52 loaves (rounded off to the nearest integer). In this way, the shop owner could decide to procure 49 loaves on Sundays, 52 on Mondays, etc.

Now, you may like to evaluate the benefits if the owner decides to make forecasts in this way, by doing the following exercise.

-
- E2) Compute the forecasts for Tuesday to Saturday as explained above. Write down the forecast errors for the three weeks and compare them with those of Table 2. What is the profit over 3 weeks according to the new decision?
-

Since you have solved E2, you know that the revised forecasting method would lead to a profit of Rs.2015/- . You also calculated the profit got by applying the previous forecasting method as Rs.1832/- . So, the method just discussed is superior to the former one.

* * *

The example above was used to give you a flavour of forecasting — how forecasting helps in decision making, what forecasting errors are and how to compute them. Generally, in practice you may not find it easy to use models to produce very accurate forecasts. This is because of two reasons :

- (i) the data may not follow a pattern that can be described by any mathematical model, and
- (ii) the pattern of the data may change all of a sudden.

However, there are ways of making reasonable forecasts. In the next two sections you will see how forecasting models are built, and various elements involved in building such models.

10.3 TIME SERIES

In the previous section you have seen examples of how data collected over a period of time can help in forecasting. You have also seen that forecasting involves studying the behaviour of a characteristic over time and examining the data for any patterns. Then the forecasts are made by assuming that the characteristic will continue to behave according to the same pattern in future. The data gathered could be sales per week, units of output per day, the cost of running a company per month, and so on.

The data on any characteristic collected with respect to time over succeeding time periods is called a time series. For example, in Table 1 we had a day-by-day time series for the demand of bread.

Some time series cover a period of several years. For example, the Andhra Pradesh government wanted to study the changes in the cropping pattern, to

be able to predict the future economic needs of the agricultural sector. For this purpose, they gathered the pertinent data, some of which are given in Table 3 below.

Table 3 : A Time Series for Crop Yield

Sl.No.	Year	Crop Area (Hectares)		Yield/Hectare (Tons)		Rainfall (Cms.)
		Rice	Sugar Cane	Rice	Sugar Cane	
1	1955	27.23	0.71	1137	7420	1064
2	1956	29.27	0.77	1163	8178	1128
3	1957	28.31	0.77	1180	8434	847
4	1958	30.10	0.74	1250	9304	1063
5	1959	30.81	0.83	1244	8605	1030
6	1960	29.61	0.91	1238	8888	851
7	1961	33.92	0.96	1239	8139	1017
8	1962	34.75	0.91	1220	9809	1134
9	1963	33.57	1.24	1292	8701	891
10	1964	34.60	1.45	1447	7477	920
11	1965	31.40	1.36	1262	8602	680
12	1966	33.23	1.08	1328	7778	948
13	1967	33.99	1.23	1375	8200	817
14	1968	28.49	1.56	1231	8180	787
15	1969	34.69	1.58	1248	7074	990
16	1970	35.21	1.20	1359	7923	956
17	1971	30.41	1.19	1551	9914	692
18	1972	29.28	1.34	1454	8245	727
19	1973	33.78	1.78	1653	8284	894
20	1974	35.53	1.95	1604	8570	848
21	1975	38.95	1.71	1657	7577	1104
22	1976	35.65	1.79	1410	7727	1024
23	1977	36.63	1.95	1565	7940	873
24	1978	39.79	1.62	1907	7067	1150
25	1979	34.69	1.39	1859	8222	743
26	1980	36.00	1.72	1991	7859	884
27	1981	38.24	2.21	2102	9142	945
28	1982	36.38	2.05	2156	7922	819
29	1983	41.63	1.72	2161	7332	1198
30	1984	34.98	1.70	2021	7322	734
31	1985	34.52	1.65	2264	7483	865
32	1986	34.59	1.68	1951	5754	868
33	1987	32.07	1.77	2258	6902	954
34	1988	42.18	1.96	2572	7690	1144
35	1989	42.06	2.10	2403	7699	1343
36	1990	40.36	2.28	2442	7281	982
37	1991	39.36	2.36	2400	7496	981
38	1992	36.04	1.71	2495	7107	837
39	1993	35.47	1.75	2759	7676	817
40	1994	36.37	2.09	2609	7150	1018
41	1995	36.92	3.58	2498	7303	971

Source : Directorate of Economics and Statistics, A.P.

Note that the successive time period in Table 3 is one year. Depending upon the characteristic of interest of the forecaster, this successive time period can be a day, a week, a month, a quarter, a year, etc.

The fluctuations you see in each time series is the net result of the effect of several forces acting on the characteristic of interest. For example, the yield of rice (or sugar cane) depends upon irrigation facilities, quality and availability of fertilizers, weather conditions, transportation facilities, etc. Scientists have classified the effects of such forces on a time series into four broad categories. They are

- i) long-term (or secular) trend,
- ii) seasonal variations,
- iii) cyclic variations,
- iv) irregular (or random) variations.

These are the four basic components of any time series on which the forecasting models are built. Let us try to understand each of these components one by one.

10.3.1 Long-term Trend

Look at the data on yield of rice in Table 3. The data are presented graphically in Fig. 2.

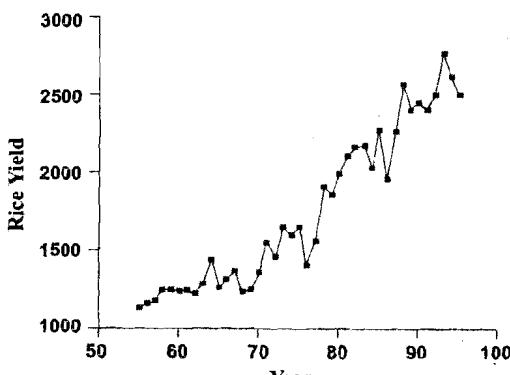
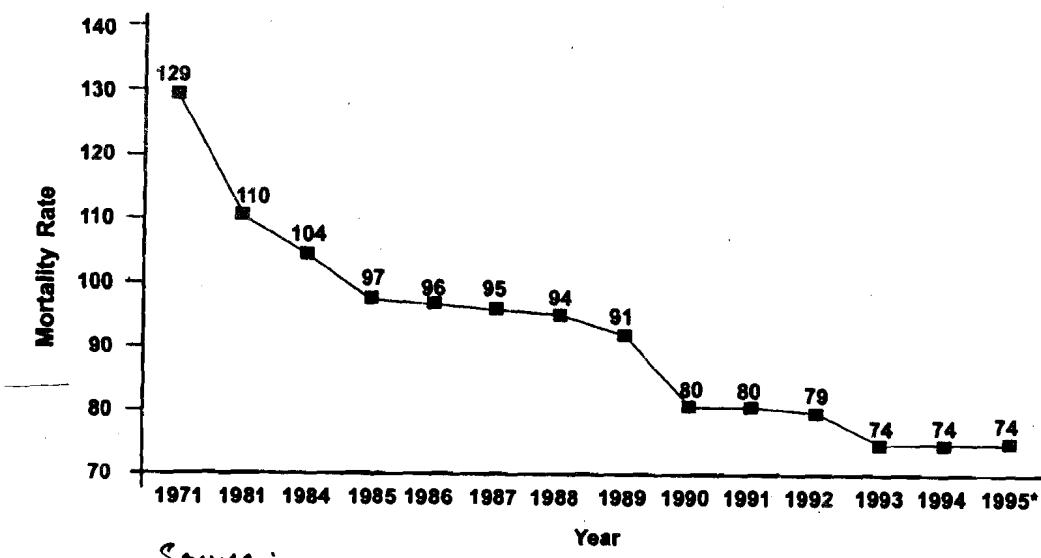


Fig. 2

From the graph, you can clearly see a general upward trend in the yield over a period of 40 years, though there are downward movements in between. The upward trend can possibly be attributed to better methods and facilities, use of new breeds of rice, etc. In fact, many business and economic statistics show upward trends over long periods.

Of course, there are series which show downward trends as well. For instance, the data on sugar cane yield presented in Table 3 exhibits a downward trend. Another example is the mortality rate of children below the age of 10 years in India (see Fig. 3). This also shows a very sharp downward trend over a long period of time.



Source :

Source : Registrar General of India

Fig. 3 : Infant mortality rate per 1,000 live births in India

There are characteristics which do not show any trend over a 15-year period or a longer period. For instance, in Fig. 4 we present the rainfall data of Table 3 graphically. Can you see any general upward or downward

trend in it? In a time series where there is no trend, the long-term trend component will be absent in the models used for forecasting such time series.

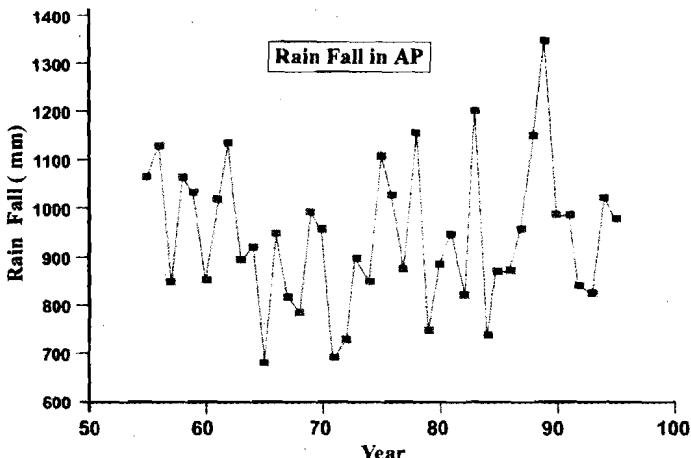


Fig.4 : Example of a time series with no trend

So, we have just seen examples of the long-term trend, which is an upward (or downward) movement in a time series over a long period of time, usually 15 years or a longer period. Here is an exercise about this now.

-
- E3) Look at your newspaper or at events around you, and give one example each of a time series with a downward trend, an upward trend and with no trend.
-

In this sub-section we introduced you to a long-term effect on a characteristic of interest. Let us now look at the second component that we had listed as an effect on the characteristic.

10.3.2 Seasonal Variations

Suppose a readymade garment's manufacturer wants to forecast the sales of cotton shirts. He studies the data he has for the period 1995-2000. This data is of quarterly (i.e., 3-monthly) sales, which are given in Table 4.

Table 4 : Quarterly Sales of Cotton Shirts

Year	Shirt Sales			
	1st quarter	2nd quarter	3rd quarter	4th quarter
1995	100	243	250	120
1996	95	250	239	113
1997	111	227	241	110
1998	107	232	230	100
1999	110	230	237	97
2000	92	245	229	98

Let us draw the graph of these sales by the quarter (see Fig. 5) for 3 years.

From Fig. 5 you can see that for each year the sales are low in the first and fourth quarters, but high, and more or less the same, in the second and third quarters. So, within a year, the pattern is different in different quarters.

However, the same pattern repeats every year. We have shown the sales for only 3 years in Fig. 5. The sales for the other years follow the same pattern. This kind of repetition of a pattern within a time period (of a year in this case), and repeated every year is an example of a seasonal variation.

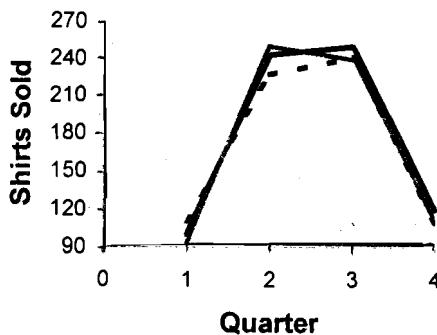


Fig. 5 : Quarterly sales of shirts

More generally, a **seasonal variation** of a characteristic is a pattern found in the data **within** a time period (a year in the example above) whose shape repeats in every successive period (in the example above, year after year). For instance, if you study the data of annual rainfall in India, but given month-wise, you will find high rainfall during the months of July and August, and almost no rainfall during April and May.

Seasonal variations need not only be due to changes in the natural weather conditions. They could be due to human-made conditions as well. For example, the number of STD phone calls made during a day will vary with the time slots specified by the telephone department (at present MTNL has two time slots with varying charges). Similarly, the number of passengers travelling in city buses is usually low on Saturdays and Sundays as compared to other (working) days of a week.

Now, try the following exercises.

- E4) Give one example each from your own experience of a time series with seasonal variations
 - i) due to natural weather conditions;
 - ii) due to human-made conditions.

- E5) Give an example of a time series which has no seasonal variation component.

As you may have realised, the seasonal variation plays an important role in planning or forecasting **only over a short period** like a year or less. However, there is a somewhat similar component of a time series that is, in a sense, linked with the long-term trend. Let us discuss this component now.

10.3.3 Cyclic Variations

When you were studying the data in Table 3, you may have noticed that the data on crop area of sugar cane presented there seems to increase and decrease repeatedly over the 41 years. In Fig.6, we present the same data graphically.

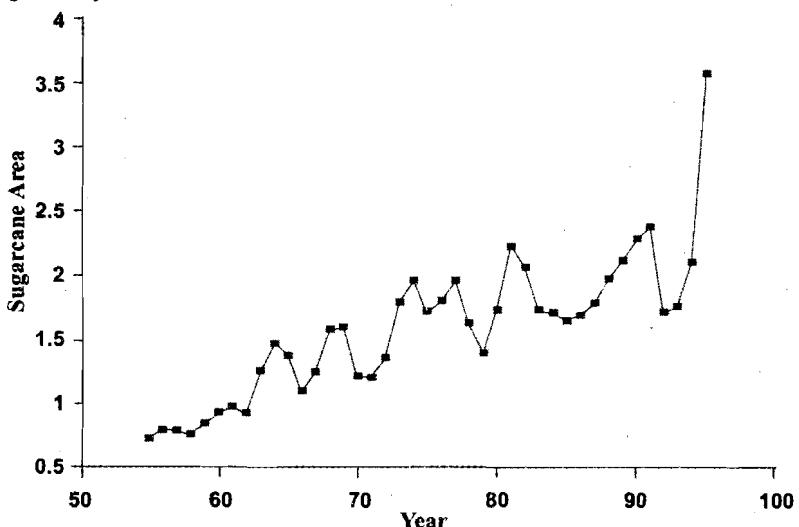


Fig. 6 : The crop area of sugar cane in Andhra Pradesh from 1955-1995

From the graph above, you can clearly see that the crop area is increasing upto a certain point and then dropping, then again increasing and dropping. This kind of movement in a time series is known as cyclic variation.

The time series from one peak to the next (or from one lowest turning point to the next lowest turning point) is known as a **cycle**. Unlike the seasonal variation, the length (or duration) of a cycle in a **cyclic variation** is not periodic, as you can see from Fig. 6. One cycle is from 1962-66, another is from 1966-71, etc., that is, they are of varying lengths.

Some time series may not exhibit any cyclic variations. Can you think of one such series? What about the time series of child mortality in India which is graphed in Fig. 3? There is no cyclic component visible here.

Cyclic variations are common among commercial and economic time series, in which the length of a cycle could vary from 2 to 10 years. From Fig. 5 and Fig. 6, you may say that seasonal variations also show peaks and troughs. But the duration of a seasonal variation is short, usually a year or less. The generally accepted convention is that a **cyclic variation is called a cycle only when its duration is more than a year and a seasonal variation otherwise**.

Why don't you try an exercise now?

- E6) Consider the time series you presented while doing E3. Examine them for cyclic variations.

Now we shall briefly consider the fourth component of a time series.

10.3.4 Irregular Variations

Suppose I am studying the trends in the male vs. female birth rates in North India. While going through the annual figures per thousand from 1973 to 2000, I find that it is following a definite pattern — the number of female births is going down steadily. This pattern is disturbed suddenly at one point (in 1997) when the number of female births suddenly rises. Then the pattern continues again.

Such unexplained variation in a time series is called **random variation**, or **irregular variation**. It is the result of one or more chance causes which are purely random and unpredictable. So, this factor is represented by a random variable. The values it takes are, in fact, the estimates of the forecast errors. Therefore, it is expected that this is i.i.d normal with mean 0 (see Unit 2). Therefore, it is usually not taken into consideration while doing long-term planning and forecasting.

Here is an exercise about such variations.

-
- E7) Give an example of a time series and an irregular variation in it.
-

So, you have seen the four basic components of any time series. Any given time series may or may not involve all four components. Here is an exercise based on this.

-
- E8) Which of the four basic components have an effect on the characteristic under consideration in the following time series data? Explain your choice of components.
- i) Number of cars produced monthwise by a leading car manufacturing company from April '91 to March '98 (there will be 84 observations in this time series).
 - ii) Number of fiction books issued by a lending library daywise from Jan. 1, 1999 to December 1, 1999.
 - iii) The yarn production data shown in Table 5.
-

As you have seen, a time series helps us to see a pattern in the long-term behaviour of the particular characteristic we are studying. You have also seen that a time series, in general, has four basic components. The question is : how are these components put together so that a forecast analyst can be helped to understand the phenomena affecting the path followed by a time series? Let us try and answer this now.

10.4 FORECASTING MODELS

In this section, the key issue we shall discuss is that of appropriate models which explain the available time series data reasonably well. We shall restrict our discussion to two commonly used forecasting models — the additive model and the multiplicative model. While doing so, we shall use the notation y_t for the value of the time series at the time t . Since all time series will be in chronological order (i.e., in order of successive time periods), we can use serial numbering for time t . For example, for the

Table 5 : Annual Yarn Production of Andhra Pradesh (in 1000 tons)

Year	Production
1971	29.9
1972	26.7
1973	25.5
1974	32.5
1975	30.1
1976	30.0
1977	28.3
1978	30.5
1979	31.6
1980	32.2

data in E8 (iii) (Table 5), we can use $t = 71, 72, \dots, 80$, in which case the time series will be $y_{71} = 29.9$, $y_{72} = 26.7$, and so on. We can also serialise this data using $t = 1, 2, \dots, 10, 1$ standing for 71, 2 for 72, and so on. Then the time series will be $y_1 = 29.9$, $y_2 = 26.7$, and so on.

So let us look at the models, one-by-one.

10.4.1 The Additive Model

One of the most widely used models is the additive forecasting model. In this model it is assumed that at any time t , the time series is the sum of all the components. Symbolically, the model is

$$y_t = T_t + C_t + S_t + I_t$$

where T_t , C_t , S_t , I_t are the long-term trend, cyclic, seasonal and irregular variations, respectively. Furthermore, it is assumed that the effect of the cyclic component (C_t) remains the same for all cycles and that the effect of any seasonal variation (S_t) remains the same during any year (or corresponding period). Similarly, it is assumed that the irregular component (I_t) has the same effect throughout. (In other words, it is assumed that I_t is i.i.d normal (see Unit 2) with mean 0.)

As you have seen in Sec.10.3, it is not necessary that every time series must include all the four components. For instance, the model for the annual rice yield data does not have a seasonal component and the model for the annual rainfall data does not have a cyclical component.

Let us consider an example of the use of the additive model.

Example 3 : Let us revisit the situation in Example 1. To fit the model we use fresh data collected during five weeks in November and December, 1998, regarding these sales. These are presented in Table 6 below.

Table 6 : Bread Loaf Sales Data

Day	Week beginning :					Average
	Nov. 1	Nov. 8	Nov. 15	Nov. 22	Nov. 29	
Sun	45	52	55	56	64	54.4
Mon	46	53	59	56	60	54.8
Tue	48	45	46	51	57	49.4
Wed	47	51	53	58	55	52.8
Thu	58	61	60	66	73	63.6
Fri	58	62	61	61	70	62.4
Sat	51	56	65	64	65	60.2
Average	50.4	54.3	57.0	58.9	63.4	56.8

Since the data are there only for 5 weeks, we shall assume that the cyclic component is absent in this time series. Therefore, our model will be

$$y_t = T_t + S_t + I_t, t = 1, 2, \dots, 35.$$

Using a common sense (**ad hoc**) approach, let us identify and measure the trend and seasonal components. Here the seasonal variation is reflected by variations within each week, while the long-term trend is reflected by the movement of weekly average sales. For convenience, we have already computed the weekly and daily averages in Table 6. Notice that the weekly

averages (last row) are showing an upward trend — from the first week to the second there is an increase of 4 units in the average, from the 2nd to the 3rd weeks 3 units, 3rd to 4th weeks 2 units, and from the 4th to 5th weeks 4 units. Therefore, we can conclude that every week there is an increase of $3.2 (\frac{4+3+2+4}{4})$ units on an average. If we extend this trend to future weeks, the sixth week average sales should be 66 loaves (again rounded off for simplicity), the seventh week's should be 69, and so on. So, looking at the average, we can now write down the trend component explicitly as follows :

$$\begin{aligned} T_t &= 50 \text{ for } t = 1, 2, \dots, 7 \text{ (1st week)} \\ &= 54 \text{ for } t = 8, \dots, 14 \text{ (2nd week)} \\ &\vdots \\ &= 66 \text{ for } t = 36, \dots, 42 \text{ (6th week)} \end{aligned}$$

All the averages are rounded off to the nearest integer for simplicity.

Let us now estimate the seasonal component. Since the seasonal variation is reflected by the variation within each week, we can estimate this by subtracting any week's average from that week's observations. For example, the seasonal variations from the first week's data are obtained by subtracting 50 (the rounded off first week's average) from that week's observations. The differences thus obtained are presented in Table 7.

Table 7 : Estimates of Seasonal Components For Data in Table 6

Day	Week					Average
	1st	2nd	3rd	4th	5th	
Sunday	-5	-2	-2	-3	1	-2
Monday	-4	-1	2	-3	-3	-2
Tuesday	-2	-9	-11	-8	-6	-7
Wednesday	-3	-3	-4	-1	-8	-4
Thursday	8	7	3	7	10	7
Friday	8	8	4	2	7	6
Saturday	1	2	8	5	2	4

Note : Averages are rounded off to the nearest integer.

Recall that we have mentioned that the effect of seasonal variations remains the same over all successive periods (the period is a week in this case). In other words, we would expect the differences in each week to be identical. However, this does not happen in practice. Therefore, the seasonal component will be estimated by averaging out the differences over the weeks. In the last column in Table 7 we have noted these averages, which give us the seasonal components. So, for example,

$$S_t = -2 \text{ for } t = 1, 8, 15, 22, 29, \dots \text{ (corresponds to Sundays)}$$

$$S_t = -7 \text{ for } t = 3, 10, 17, \dots \text{ (corresponds to Tuesdays)}$$

It is now your turn to write down the values of S_t for the other days of the week.

- E9) Write down the values of S_t for $t = 2, 4, 6, 16$ and 26 in the example above.

We use \hat{y} to denote the predicted value of y .

What remains to be estimated is the irregular component, I_t . This component can be estimated by subtraction (i.e., using $I_t = y_t - T_t - S_t$) for the available data. However, estimating specific values of I_t usually does not interest the forecaster. What will interest her is analysing the distribution of the values of I_t because this will help in providing confidence intervals to the forecasts. In fact, the values of I_t are nothing but the estimates of forecast errors (recall that you have come across forecast errors in Table 2, E1 and E2). Such errors are expected to have a zero mean, that is, they cancel each other out in the long run. So, **future forecasts are made by adding only the non-random components**. For example, the forecast of y_{36} is given by

$$\hat{y}_{36} = T_{36} + S_{36} = 66 + (-2) = 64.$$

* * *

Why don't you try a related exercise now?

- E10) Using the additive model for finding the trend in the bread sales, forecast the 6th, 7th and 8th weeks sales. Also compare them with the actual sales given in Table 8 below.

Table 8 : Actual Bread Sales

Day Week	Sun	Mon	Tue	Wed	Thurs	Fri	Sat
6	59	65	58	62	79	66	69
7	64	66	60	57	81	70	69
8	69	75	67	69	78	74	73

Let us now consider another model used for forecasting.

10.4.2 The Multiplicative Model

We briefly outline this method, which is really a multiplicative version of the earlier one. In the additive model, we have assumed that the time series is **the sum** of the trend, cyclical, seasonal and random components. From practical experience, scientists have found that additive models are appropriate when the seasonal variations remain unchanged (that is, the seasonal variations do not depend on the trend of the time series). However, in practice, there are a number of situations where the seasonal variations change over time, as you will see in Example 4 below. When the seasonal variations exhibit an increasing or decreasing trend, we can try the **multiplicative model**. In the multiplicative model it is assumed that the time series is obtained as a **product** of the four time series components, that is,

$$y_t = T_t \cdot C_t \cdot S_t \cdot I_t.$$

Multiplicative models are found to be appropriate for many economic time series data such as data related to production of electricity, number of passengers going abroad, consumption of cold drinks, etc. In the following example, we will briefly describe the application of this model.

Example 4 : Examine the data on coconut sales in Hyderabad from 1995 given in Table 9. In each year, the number of coconuts sold are recorded for three seasons: (i) Season I – March to June, (ii) Season II – July to October, and (iii) Season III – November to February.

Table 9 : Coconut Sales Data

Period (t)	Year	Season	Coconuts sold (y_t) (in lakhs)	Period (t)	Year	Season	Coconuts sold (y_t) (in lakhs)
1	1994-95	I	14	12	1997-98	III	90
2	1994-95	II	35	13	1998-99	I	22
3	1994-95	III	65	14	1998-99	II	50
4	1995-96	I	14	15	1998-99	III	105
5	1995-96	II	43	16	1999-00	I	23
6	1995-96	III	77	17	1999-00	II	30
7	1996-97	I	16	18	1999-00	III	120
8	1996-97	II	40	19	2000-01	I	27
9	1996-97	III	84	20	2000-01	II	68
10	1997-98	I	17	21	2000-01	III	132
11	1997-98	II	46				

First let us observe the time series plot. This is given in Fig. 7.

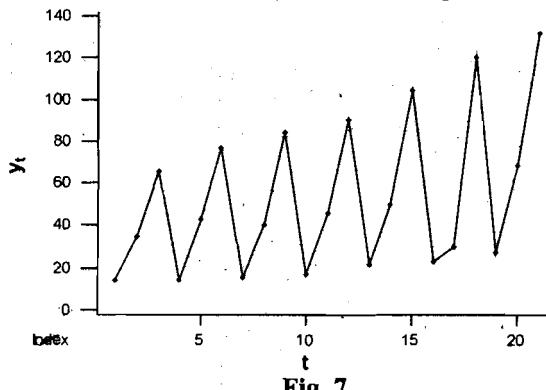


Fig. 7

From the plot you can easily see two things :

- (i) the sales are gradually increasing (this indicates the increasing trend), and
- (ii) the seasonal variation clearly exists, and, more importantly, it is increasing with the increasing trend.

Here is an exercise for you about this.

- E11) Find out the **seasonal range** in each year. (e.g., the seasonal range for the year 1994-95 is 51 (= 65 – 14). Do you find any trend in the seasonal range over the years?

While doing E11, you would have realised that the seasonal variation is increasing with an increasing trend. So, as we said at the beginning of this sub-section, we should try the multiplicative model in this case. The cyclic variation that you see in Fig. 7 is actually seasonal variation. So, we drop

the cyclic component C_t , and include the seasonal component S_t in our model. Consequently, our model will be :
 $y_t = T_t S_t I_t$.

We will see how we can estimate the time series components T_t , S_t and I_t , using formal methods, in the next section.

* * *

Note that in order to apply the multiplicative model the time series should have positive values. So, if we wish to use the multiplicative model to understand a time series with negative values, then we need to convert the time series to positive values by adding a suitable constant to each entry.

In this section you have seen examples of building forecasting models in some cases. Our approach was an ad hoc one, based on common sense. However, this is not the way analysts do it. In the next section you will learn some scientific methods of analysing time series data.

10.5 FORECASTING LONG-TERM TREND

Here we use the method of least squares, the moving average method and the method of exponential smoothing for finding the component T_t , mentioned in the models above. As you will see, identifying the trend in a time series requires elimination of other components from the time series.

Let us start with the simplest of these methods, which you studied in Unit-9.

10.5.1 The Least Squares Method

Let us once again look at the bread sales data presented in Table 6. The same data are reorganised in Table 10 day-wise.

Table 10.: Bread Sales Data

Day (t)	y_t	Day (t)	y_t
1	45	19	60
2	46	20	61
3	48	21	65
4	47	22	56
5	58	23	56
6	58	24	51
7	51	25	58
8	52	26	66
9	53	27	61
10	45	28	64
11	51	29	64
12	61	30	60
13	62	31	57
14	56	32	55
15	55	33	73
16	59	34	70
17	46	35	65
18	53		

Now, what do you expect its trend component to be like? Does the trend component of the time series increase (or decrease) at a constant rate? If it does, then the time series is said to have a **linear trend**, that is, it is a linear function of time. What this means algebraically is that if y_t has a linear trend, then we would expect $T_t = a + bt$, where a and b are constants. Recall (from Unit 9) how you fit linear equations. We can use the method of least squares. Here T_t is our dependent variable and t is our independent variable. Setting $y = T_t$ and $x = t$, we get

$$\bar{x} = 18, \bar{y} = 56.8, S_{xx} = 3570, S_{xy} = 1732, S_{yy} = 1715.6.$$

Therefore, the parameters a and b of the best fit linear equation are estimated as

$$\hat{b} = S_{xy}/S_{xx} = 0.485, \text{ and}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 48.06.$$

So, the regression equation in this case is given by

$$T_t = 48.06 + 0.485t. \quad (1)$$

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

This equation can now be used to obtain the trend component T_t .

If you work out the square of the correlations coefficient (see Sec. 9.3.2), you will find $R^2 = 0.49$. From Unit 9, you know that this means that this regression is reasonably reliable, but could be better.

Notice that in our earlier approach in Example 3, we had the same value for the trend component during any day of the week. In this regression approach, we get different values of the trend even within a week. In the following exercises we ask you to compare some of these different trend values.

- E12) Compute T_t for our bread sales example for Wednesdays of each of the five weeks mentioned in Table 6, using Equation (1) above.
Compare them with what we have got in Example 3.
- E13) Fit the linear trend using part of the sugar cane crop area data presented in Table 11 below (extracted from Table 3).

Table 11 : Sugar Cane Crop Area in AP (in lakh hectares)

Year	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
Area	0.96	0.91	1.24	1.45	1.36	1.08	1.23	1.56	1.58	1.20

Not all data exhibit a linear trend. Sometimes, by looking at the data points, it becomes reasonably clear that the time series data exhibit non-linear trends. For example, if you examine the rice yield data presented in Fig.2, there is a clear indication of a non-linear trend.

Let us look at another set of data, the population of Andhra Pradesh presented in Table 12 below.

Table 12 : Andhra Pradesh Population

Census	Population (in 10000s)	Census	Population (in 10000s)
1901	1906	1951	3111
1911	2144	1961	3598
1921	2142	1971	4350
1931	2420	1981	5355
1941	2729	1991	6651

Courtesy : Directorate of Economics and Statistics, AP

Look at the graph of the time series above given in Fig. 8 below. The points are certainly not lying around any line. Some non-linear curve may fit the data very well.

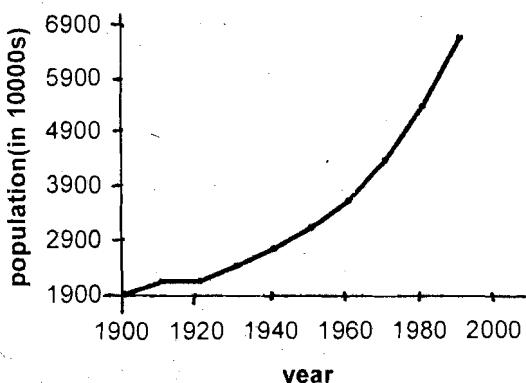


Fig. 8 : Graph of data in Table 12

A number of standard forms of curves have been found to be useful for fitting the data, in practice. These are polynomial curves, exponential curves and growth curves. The graphical plots of the time series is useful in identifying the form of the trend curve. Let us illustrate this with an example.

Example 5 : Let us try to fit a second degree equation to the trend of rice yield data of Table 3 using the method of least squares. We first plot the time series (see 'actual' curve in Fig. 9 below). Looking at it, it seems that a line will not be the best fit. Let us try a quadratic polynomial. The equation would be

$$T_t = a_0 + a_1 t + a_2 t^2, \quad t = 55, 56, \dots, 95.$$

The constants a_0 , a_1 and a_2 are estimated by fitting a multiple linear regression model with two independent variables, $x_1 = t$ and $x_2 = t^2$, and the dependent variable $y = T_t$. Refer to Sec. 9.4.2 to recall the method of computing \hat{a}_0 , \hat{a}_1 and \hat{a}_2 , the estimates of a_0 , a_1 and a_2 , respectively.

The first step in this is to tabulate the values of (x_1, x_2, y) . There will be 41 observations of (x_1, x_2, y) . From Table 3 we get the first one as $(55, 55^2, 1137)$, the second as $(56, 56^2, 1163)$, and so on. Next, we have to compute S_{11} , S_{22} , S_{12} , S_{1y} and S_{2y} . Then, \hat{a}_1 and \hat{a}_2 are obtained by solving the normal equations (10) and (11) of Sec. 9.4.2. Finally \hat{a}_0 is obtained from Equation (12) of Sec. 9.4.2. We shall omit the details of this calculation

Notice that we are using year number for t rather than the serial number.

The normal equations are derived from the least squares principle.

and present the final results directly. The estimates are given by $\hat{a}_0 = 2660$, $\hat{a}_1 = -67.1$ and $\hat{a}_2 = 0.716$. Therefore, the trend curve is given by

$$T_t = 2660 - 67.1t + 0.716t^2 \quad (2)$$

Here $R^2 = 0.94$ and the standard error $s = 126.8$.

The trend curve fitted in (2) and the actual time series values are plotted in Fig. 9. This may give you an idea about how good the fit is.

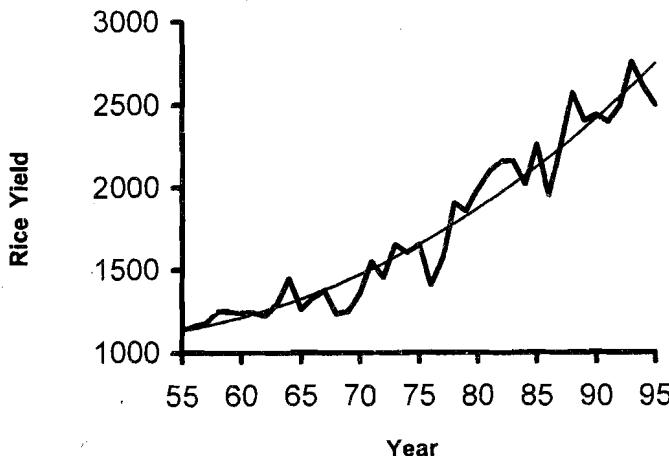


Fig. 9 : A quadratic fit to the trend curve

From Fig. 9 we see that the quadratic polynomial is a good fit for the trend of the rice yield data.

* * *

Now you should try to see how a second degree equation fits the population data of Table 12.

- E14) Fit the trend curve to the population data of Table 12 using a second degree equation, i.e., $T_t = a_0 + a_1 t + a_2 t^2$. Use $t = 1, 2, \dots, 10$. How good is your fit?

Let us now consider a situation where a non-linear trend is best fitted by an exponential curve.

Example 6 : Our country's population has been growing at a tremendous rate. In such a situation, when the growth rate is increasing fast, an exponential trend curve is probably the best fit to the data.

Examples of equations representing exponential curves are $y = 2(3^t)$ and $y = (0.7)(2^t)$.

In general, it is of the form

$y_t = ab^t$, where a and b are positive constants.

Supposing we wish to fit this curve to the population data of Table 12. How do we go about it? Our equation for the trend curve would be

$$T_t = ab^t, \quad (3)$$

where a and b are positive constants.

Notice that if we take the common logarithm (to the base 10) on both sides of Equation (3), we get

$$\log T_t = \log a + t \cdot \log b \quad (4)$$

If we put $z_t = \log y_t$, $Z_t = \log(T_t)$, $a_0 = \log a$ and $a_1 = \log b$, we can rewrite (4) as

$$Z_t = a_0 + a_1 t.$$

Now you know how to find out the values of a_0 and a_1 by the least squares method. Once you obtain the estimates of a_0 and a_1 , you can calculate the values of a and b using antilogarithms. The calculations are done for you in Table 13 below.

Table 13 : AP Population Values For Curve Fitting

Census	t	Population (in 10,000s) (y _t)	$z_t = \log y_t$	Forecast for z_t (Z_t)	Forecast for y_t (T_t)
1901	1	1906	3.2801	3.2333	1711
1911	2	2144	3.3312	3.2925	1961
1921	3	2142	3.3308	3.3517	2248
1931	4	2420	3.3838	3.4109	2576
1941	5	2729	3.4360	3.4701	2952
1951	6	3111	3.4929	3.5293	3383
1961	7	3598	3.5561	3.5885	3877
1971	8	4350	3.6385	3.6477	4443
1981	9	5355	3.7288	3.7069	5092
1991	10	6651	3.8229	3.7661	5836

Z_t is the forecast for $\log(y_t)$ and is given by $Z_t = 3.1741 + 0.0592 \times t$, $R^2 = 0.96$, $S_e = 0.039$.

The original fit is obtained by taking antilogarithms. This is given by $T_t = \text{antilog}(3.1741)(\text{antilog } 0.0592)^t = 1493 \times (1.1462)^t$.

* * *

Why don't you try an exercise now?

- E15) (a) Compare the exponential fit with the quadratic fit you obtained in E14.
 (b) The actual population of AP in 2001 is 7,57,27,541. Which of the two curves above gives a better forecast? What will the population be in 2011 according to both the curves fitted to the data?

- E16) Estimate T_t using regression for the data in Example 4.

So far we have looked at two methods of forecasting trends. There is another method based on finding the averages of the data. Let us look at its strong and weak points.

10.5.2 The Method of Moving Averages

This method aims at identifying the long-term trend by eliminating seasonal variations. While doing this, the method also indicates the presence of

seasonal and cyclic variations, if any. To appreciate this, let us apply this method to our bread sales data presented in Table 6.

Example 1 (Contd.) : To apply the method we need to present the data in a single column (in chronological order). This is done in Column 2 of Table 14 below.

Table 14 : Moving Averages for the Bread Sales Data

Day (t)	Sales(y _t)	Moving averages with length :				
		4	5	7	8	10
1	45					
2	46	46.50				
3	48	49.75	48.80			
4	47	52.75	51.40	50.43	50.63	
5	58	53.50	52.40	51.43	51.63	
6	58	54.75	53.20	52.43	51.50	50.30
7	51	53.50	54.40	52.00	51.88	50.90
8	52	50.25	51.80	52.57	53.63	52.40
9	53	50.25	50.40	53.00	54.13	53.80
10	45	52.50	52.40	53.57	53.88	54.70
11	51	54.40	54.29	54.38	54.40	
12	61	55.00	54.71	55.25	54.50	
13	62	57.50	57.00	55.57	54.38	54.00
14	56	58.50	58.60	55.71	55.38	54.10
15	55	58.00	55.60	56.00	56.50	54.80
16	59	54.00	53.80	55.86	56.50	56.40
17	46	53.25	54.60	55.71	56.88	57.80
18	53	54.50	55.80	57.00	56.88	57.30
19	60	55.00	57.00	57.14	57.00	56.70
20	61	59.75	59.00	56.71	56.00	56.20
21	65	60.50	59.60	57.43	57.50	56.50
22	56	59.50	57.80	58.14	59.13	57.20
23	56	57.00	57.20	59.00	59.25	58.70
24	51	55.25	57.40	59.00	59.63	59.80
25	58	57.75	58.40	58.86	59.50	60.20
26	66	59.00	60.00	60.00	60.00	60.10
27	61	62.25	62.60	60.57	60.13	59.30
28	64	63.75	63.00	61.43	60.63	59.20
29	64	62.25	61.20	61.00	62.50	60.90
30	60	61.25	60.00	62.00	63.00	62.80
31	57	59.00	61.80	63.29	63.50	63.50
32	55	61.25	63.00	63.43		
33	73	63.75	64.00	63.33		
34	70	65.75				
35	65					

For the time being ignore Columns (3), (4), (6) and (7) in the table. Let us see how we have obtained the entries in Column (5). We first compute the average of the first seven observations in Column (2), i.e., $\frac{45+\dots+51}{7} =$

50.43. We place this figure in Column (5) in line with Day 4, the mean of Days 1 to 7. Next we compute the average of the seven observations y_2, y_3, \dots, y_8 , (that is, we drop y_1 and include y_8). This is 51.43. We place it in Column (5) in line with Day 5. In this way we compute the averages of seven consecutive observations, each time dropping the first observation and including the next observation. In this way, the last entry in Column (5), 63.33, is the average of the last seven observations of Column (2), and is placed in line with Day 32. These averages we just computed are called the moving averages of length 7.

The averages are called 'moving' averages because at each stage of calculating averages we move from one period to the next period.

Now consider Column (4). This consists of the moving averages of length 5. So, the first entry is $\frac{y_1+y_2+y_3+y_4+y_5}{5} = 48.80$. This figure is written

in line with Day 3. The next entry in Column (4) is $\frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}$, and so on.

We have just seen how to find moving averages of odd length (7 and 5, respectively). Let us now see how to compute the moving averages of even length. Suppose we wish to compute the moving averages of length 4 for the bread sales data. We compute the average of the first 4 observations of Column (2) in Table 14. This average is equal to 46.5. What day does it correspond to? We imagine that it corresponds to 'Day 2.5' (i.e.,

$\frac{1+2+3+4}{4}$), and place it in Column (3) at the point between the rows

corresponding to Day 2 and Day 3. Next, we compute the average of the 2nd, 3rd, 4th and 5th observations. This is 49.75. Correspond this with 'Day 3.5' in Column (3). Since 46.5 corresponds to Day 2.5 and 49.75 corresponds to Day 3.5, the average of 46.5 and 49.75 = $(46.5 + 49.75)/2 = 48.125$ should correspond to Day 3 (= average of 2.5 and 3.5).

In this way we continue, the last entry in Column (3) being $\frac{55+73+70+65}{4} = 65.75$. This is placed in line with 'Day 33.5'.

You should calculate and check for yourself that all the entries shown in Columns (3) to (7) are correct.

To make sure that you have got the definition and computation of moving averages correctly, do the following exercises.

- E17) Compute the first three moving averages of length 3 for the bread sales data and place them in line with the corresponding day numbers.
- E18) For the bread sales data, compute the 2nd and last moving averages of length 6. What days do these averages correspond to?

Now let us see the graphic representation of the moving averages. In Fig. 10, we have plotted the moving averages (of length 7) computed in Column (5), Table 14, against the corresponding day numbers. The straight line trend obtained in Sec. 10.5.1 (Equation (1)) is also drawn in the figure.

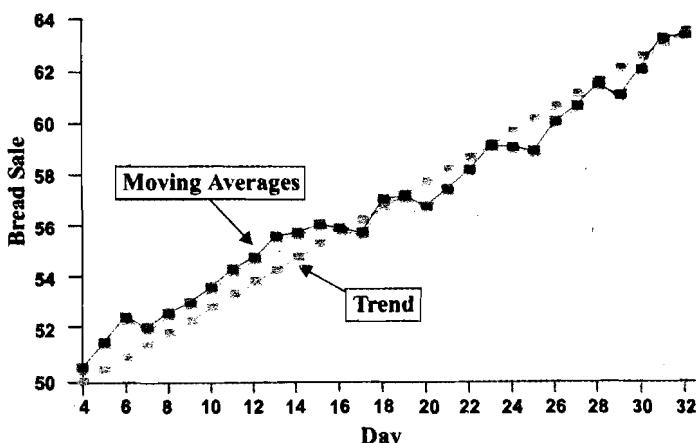


Fig. 10 : Plot of moving averages and linear trend for bread sales data

Now look at Fig. 11, where the moving averages of lengths 4, 8 and 10 (shown in Table 14) are plotted along with the trend line.

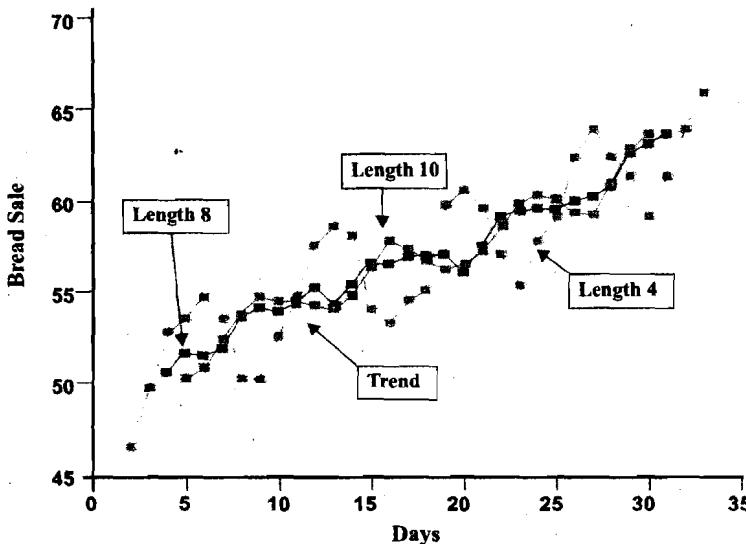


Fig. 11 : Moving averages of different lengths

You can see from Fig. 10 and Fig. 11 that fluctuations, i.e., peaks and troughs, are smallest when the length of the moving averages is 7. This is not surprising as the bread sales data have the seasonal component, the season being the **seven days of the week**. Observe from Fig. 10 that approximately half of the moving averages are above the trend line and the rest are below or on the trend line. This is to be expected whenever the linear trend is the best fit, since it represents the 'average' in a sense.

If you go back to Fig. 1, you will notice that the time series shown in it doesn't seem to show a linear, quadratic or exponential trend. In fact, it is a very jerky graph. But, related to the same data, Fig. 11 shows us that the moving averages of different lengths of the same data gives us smoother curves that show the long-term movements of the series. In this way the moving averages can be used to estimate the trends, if the time series does not help us in doing so.

* * *

From the example above, can you see what the basic idea is of using moving averages? If the time series of the data contains certain seasonal or cyclical variations, the effect of these variations can be eliminated by taking a moving average where the time period in the average equals the period of the season or the cycle. By smoothening the curve in this way, the trend doesn't get affected. For instance, in the example above the cycle is of length 7 and so the moving averages of length 7 will help us study the long-term trend, since the S_t -component of y_t gets removed.

Now, why don't you try the following exercise?

- E19) Find the moving averages of lengths 3, 4 and 5 for the rice yield data given in Table 3. Using which of these averages can the effect of the cycles be eliminated? From this exercise, what conclusions can you draw regarding the C and S components of the time series?
- E20) Using moving averages of length 3, estimate T_t for the data in Example 4.

Before we wind up this sub-section on the method of moving averages (MA), let us note the following points :

- 1) If a time series is a purely random sequence of numbers, you will find that a moving average of this time series will tend to show cyclical fluctuations. This is because a moving average is serially correlated. So, remember that many apparent cycles in moving averages may be spurious.
- 2) The peaks and troughs in the moving average may occur at different times than the peaks and troughs in the original time series (see Fig. 10 and Fig. 11).
- 3) A moving average cannot be calculated for the latest or earliest years in a time series, since the average depends on numbers that precede or occur after these years in the time series.
- 4) The method of moving averages is useful when the trend in time series data is linear, or approximately so. This is because if you compute the moving averages for such a series, given by, say, $y_t = a + bt$, you will find that they will coincide with the time series values. In fact, this is the case irrespective of the length of the moving average. However, we can't claim this for general time series which is not linear. (You may like to check this for the quadratic time series $y_t = 2 + t^2$.)
- 5) It is effective when the fluctuations in time series are regular and periodic provided the length of the moving averages is that of the period.
- 6) The method is not very useful if it has to be used for forecasting future trend values. It is basically a tool to identify trend and cyclic components in a time series. For instance, examine the data in Table 14 on bread sales. We have 35 observations here. Let us say we want to predict the 36th observation using MA. To know the MA against 36, let us say of length 3, we need to have observations for the time points 36 and 37. As we don't have these observations when we are at time 35, we can't compute the required MA. For this reason MA is not useful for forecasting.

As you have seen, time series often have so many irregular fluctuations that we aren't able to see the general trend. The method of moving averages helps in removing such fluctuations to a large extent. Therefore, we call such a method a **smoothing technique**. We shall now consider another method for smoothening the long-term trend curves.

10.5.3 Exponential Smoothing

In this sub-section we shall introduce you to another smoothing technique in which **weighted averages** are calculated. In the method of moving averages we also attach weights, equal weights to each observation that is considered. In the method we shall now discuss, the weights assigned to past and current values of the time series are different fixed positive numbers. This method is called **exponential smoothing**. Let us see why through an example.

Example 7 : Consider the rainfall data presented in Table 3. From Fig.4, it appears that there is no trend in the time series. This is the same as saying that if $T_t = a + bt$ is the trend, then $b = 0$. Therefore, if we fit a linear regression to this time series data with $b = 0$, then the least

square estimate of a is given by $\hat{a} = \frac{\sum_{t=1}^{41} y_t}{41}$, the simple average of the time series values.

Notice that the method of least squares gives equal weightage to all the observations in the time series.

Also, notice that since our model is $T_t = a + (\text{forecast error})$, our forecast for rainfall is \hat{a} for all t . This seems somewhat unreasonable because, even though there is no trend, the constant does not seem to be the same at all times t , at least going by Fig.4. The value of a seems to be higher during the period 1955 to 1965 compared to the value of a for the period 1965 to 1975. Therefore, it is logical to assume that the value of a is gradually changing over time. Therefore, we should denote it by a_t rather than a .

Accordingly we select a single weight w (called the **exponential smoothing constant**), where w lies between 0 and 1. Then we compute an exponentially smoothed series a_t as follows :

$$\begin{aligned} a_1 &= y_1 \\ a_2 &= wy_2 + (1 - w)a_1 \\ a_3 &= wy_3 + (1 - w)a_2 \\ &\vdots \\ a_t &= wy_t + (1 - w)a_{t-1} \end{aligned} \tag{6}$$

w is chosen between 0 and 1 because each a_t is a convex combination of y_t and a_{t-1} .

Two things needs to be specified here :

- i) the quantity w is a constant known as **smoothing constant**, and there is a method of choosing this constant (which we shall discuss a little later); and
- ii) we should know the value of a_0 (i.e., when $t = 1$, $a_{t-1} = a_0$) to initialise the computation of a_1, a_2, \dots . We will take this initial value of a_0 as \hat{a} , the simple arithmetic mean of all the time series values.

Let us now compute some of these quantities for the rainfall data. Firstly,

$$a_0 = \hat{a} = \frac{\sum y_t}{41} = 939.95, \text{ rounded off to the } 2^{\text{nd}} \text{ decimal place.}$$

The value of the smoothing constant w can be anything between 0 and 1. However, from experience statisticians have found that one should choose an 'appropriate' value of w between 0.01 and 0.3. Let us take, to start with, $w = 0.02$. Then, from Equation (6) we get

$$\hat{a}_1 = 0.02y_1 + 0.98a_0 = (0.02 \times 1064) + (0.98 \times 939.95) = 942.43.$$

Our forecast for y_1 at time zero is \hat{a}_0 . Therefore, the forecast error is

$$e_1 = y_1 - \hat{a}_0 = 1064 - 939.95 = 124.05.$$

The forecast error at $t = 2$ is

$$e_2 = y_2 - \hat{a}_1 = 1128 - 942.43 = 185.57.$$

Next,

$$\hat{a}_2 = 0.02y_2 + 0.98\hat{a}_1 = 0.02 \times 1128 + 0.98 \times 942.43 = 946.14,$$

and the forecast error

$$e_3 = y_3 - \hat{a}_2 = 847 - 946.14 = -99.14.$$

Now try the following exercise.

- E21) Compute \hat{a}_3 , \hat{a}_4 , the corresponding forecasts and the forecasting errors.

In the table below we have given all the forecasts and errors.

Table 15 : Exponential Smoothing of Rainfall Data

Year	Rainfall y_t	Forecast a_t	Error $e_t = y_t - \hat{a}_t$
1955	1064	939.95	124.05
1956	1128	942.431	185.569
1957	847	946.1424	-99.14238
1958	1063	944.1595	118.8405
1959	1030	946.5363	83.46366
1960	851	948.2056	-97.20561
1961	1017	946.2615	70.7385
1962	1134	947.6763	186.3237
1963	891	951.4027	-60.40275
1964	920	950.1947	-30.19469
1965	680	949.5908	-269.5908
1966	948	944.199	3.801018
1967	817	944.275	-127.275
1968	787	941.7295	-154.7295
1969	990	938.6349	51.36509
1970	956	939.6622	16.33779
1971	692	939.989	-247.989
1972	727	935.0292	-208.0292
1973	894	930.8686	-36.86861
1974	848	930.1312	-82.13123
1975	1104	928.4886	175.5114
1976	1024	931.9988	92.00116
1977	873	933.8389	-60.83886
1978	1150	932.6221	217.3779
1979	743	936.9696	-193.9696
1980	884	933.0902	-49.09025
1981	945	932.1084	12.89156
1982	819	932.3663	-113.3663
1983	1198	930.0989	267.9011
1984	734	935.457	-201.457
1985	865	931.4278	-66.42783
1986	868	930.0993	-62.09927
1987	954	928.8573	25.14271
1988	1144	929.3601	214.6399
1989	1343	933.6529	409.3471
1990	982	941.8399	40.16012
1991	981	9942.6431	48.35692
1992	837	943.4102	-106.4102
1993	817	941.282	-124.282
1994	1018	938.7964	79.20362
1995	971	940.3805	30.61955

Now look at the graphs of the time series before and after smoothing, in Fig.12. (The 'before smoothing' graph is also given in Fig.4.) You can see that the peaks are barely there in the graph of the time series after smoothing, which is the dotted curve.

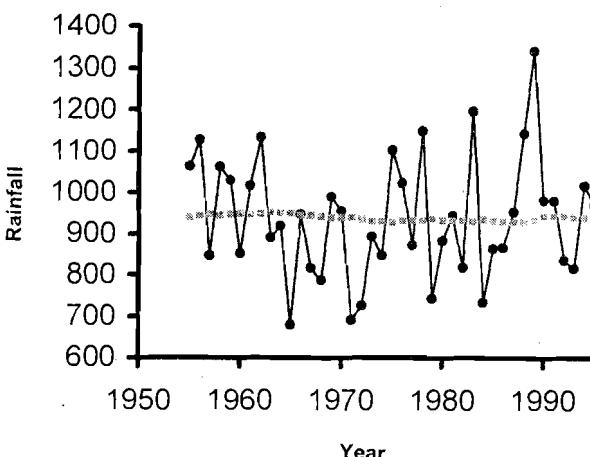


Fig.12

You can also try forecasts with other values of w , and see the curves you get.

We still haven't seen what the appropriate choice of w is. For this, we need to consider $S_E = \sqrt{\frac{e_1^2 + \dots + e_n^2}{n}}$, that is, the square root of the average error sum of squares. Since this depends on the choice of w , we shall denote it by $S_E(w)$. If we compute this quantity for our rainfall data, we get

$$S_E(0.02) = \sqrt{\frac{(124.05)^2 + \dots + (30.62)^2}{41}} = 146.36, \text{ by using the calculations shown in Table 15.}$$

We are now in a position to find out what value of w is most appropriate for our rainfall data — we should find the value of w for which $S_E(w)$ is minimum. How do we find this out? Note that if $S_E(w)$ is the minimum for $w = \alpha$, say, then $41(S_E(\alpha))^2$ will be the minimum of all values of $41(S_E(w))^2$ calculated for different values of w . So, to find the value of w that gives the least value of $S_E(w)$ it is enough to find the error sum of squares, say, $(E(w))^2$. In Table 16 we give $[E(w)]^2$ for different values of w . Note that $S_E(w)$ is small when $w = 0.0005$ (or $w = 0.001$). **The small value of w indicates that the average level of the time series is not changing much over time.**

Using the recursive equation (6) it can be shown that

$$\hat{a}_t = w y_t + w(1-w)y_{t-1} + w(1-w)^2 y_{t-2} + \dots + w(1-w)^{t-1} y_1 + (1-w)^t \hat{a}_0$$

So, for example, we get

$$\hat{y}_{31} = \hat{a}_{30} = w y_{30} + w(1-w)y_{29} + \dots + w(1-w)^{29} y_1 + (1-w)^{30} \hat{a}_0$$

Since $(1-w), (1-w)^2, \dots$ are decreasing exponentially, the weightages given to more recent observations is more. So, **all** the observations are being given weightage here, and the latest observation is being given

Table 16: Values of $S_E(w)$ for the Rainfall Data

S.No.	w	$[E(w)]^2$
1	0.0005	8,62,707
2	0.001	8,62,710
3	0.005	8,62,795
4	0.01	8,63,036
5	0.05	8,68,546

greatest weightage. In this way, this method is a refinement of the method of moving averages.

Here is a related exercise.

- E22) Apply the simple exponential smoothing procedure to the following data. Take $w = 0.1$.

Table 17 : No. of New Bank Branches Opened in a Town

Year	1981	1982	1983	1984	1985	1986	1987
No. of Branches	5	3	3	4	3	6	4

Let us now consider another **exponential smoothing forecasting method**. This is due to C.C. Holt, suggested by him in 1958.

Holt's method : We shall illustrate this procedure with the rice yield data. Suppose that the trend shown by the rice yield data is linear, namely, $y_t = a + bt + e_t$, where e_t is the error.

As before, we shall assume that a and b change gradually over time. Therefore, we denote the values of a and b at time t by \hat{a}_t and \hat{b}_t , respectively. As in simple exponential smoothing, \hat{a}_t and \hat{b}_t are smoothed using two smoothing constants w_1 and w_2 (both between 0 and 1). The recursive equations for computing these quantities and the forecasts are given by :

$$\hat{a}_t = w_1 y_t + (1 - w_1) [\hat{a}_{t-1} + \hat{b}_{t-1}], \quad (7)$$

$$\hat{b}_t = w_2 [\hat{a}_t - \hat{a}_{t-1}] + (1 - w_2) \hat{b}_{t-1} \quad (8)$$

and the forecast for the immediate future y_{t+1} is given by

$$\hat{y}_{t+1} = \hat{a}_t + \hat{b}_t. \quad (9)$$

Here too, we need the initial values a_0 and b_0 . These, together with w_1 and w_2 , should be chosen so that the sum of the squares of the forecasting errors is minimised. One suggestion to obtain a_0 , b_0 , w_1 and w_2 is to first obtain a_0 and b_0 by fitting a linear regression to one half of the time series data (a_0 as the intercept and b_0 as coefficient). Then, using them as initial values, we should obtain the values of w_1 and w_2 which minimise S_E , the square root of the average of error sum of squares (see Table 18 below). Once w_1 and w_2 are obtained, then we can change the initial values a_0 and b_0 to the intercept and coefficient of the regression line fitted to the entire time series data. However, there is no guarantee that this would lead to the best choice of a_0 , b_0 , w_1 and w_2 . In fact, this procedure of obtaining a_0 , b_0 , w_1 and w_2 in our example results in a very large value of S_E . A better choice is

$$\hat{a}_0 = 1090, \hat{b}_0 = 32, w_1 = 0.4, \text{ and } w_2 = 0.01.$$

The resulting $S_E = 133.16$. The computations are shown in Table 18. The values in the columns have been rounded off to the nearest integer for simplification in calculations.

Table 18 : Holt's Exponential Smoothing of Rice Yield Data

Year	t	Rice Yield	\hat{a}_t	\hat{b}_t	Forecast	Error
1955	1	1137	1128	32	1122	15
1956	2	1163	1161	32	1160	3
1957	3	1180	1188	32	1193	-13
1958	4	1250	1232	32	1220	30
1959	5	1244	1256	32	1264	-20
1960	6	1238	1268	32	1288	-50
1961	7	1239	1276	32	1300	-61
1962	8	1220	1272	32	1307	-87
1963	9	1292	1299	31	1304	-12
1964	10	1447	1377	31	1330	117
1965	11	1262	1350	32	1409	-147
1966	12	1328	1360	31	1381	-53
1967	13	1375	1384	31	1391	-16
1968	14	1231	1342	31	1415	-184
1969	15	1248	1322	30	1372	-124
1970	16	1359	1355	30	1352	7
1971	17	1551	1451	30	1384	167
1972	18	1454	1470	30	1481	-27
1973	19	1653	1562	31	1501	152
1974	20	1604	1597	31	1592	12
1975	21	1657	1639	31	1628	29
1976	22	1410	1566	30	1670	-260
1977	23	1565	1584	30	1596	-31
1978	24	1907	1731	31	1613	294
1979	25	1859.	1801	31	1762	97
1980	26	1991	1896	32	1832	159
1981	27	2102	1997	33	1928	174
1982	28	2156	2080	33	2030	126
1983	29	2161	2133	33	2114	47
1984	30	2021	2108	33	2166	-145
1985	31	2264	2190	33	2141	123
1986	32	1951	2114	32	2223	-272
1987	33	2258	2191	33	2147	111
1988	34	2572	2363	34	2224	348
1989	35	2403	2399	34	2397	6
1990	36	2442	2437	34	2434	8

The forecasts and the actual time series values are plotted in Fig. 13.

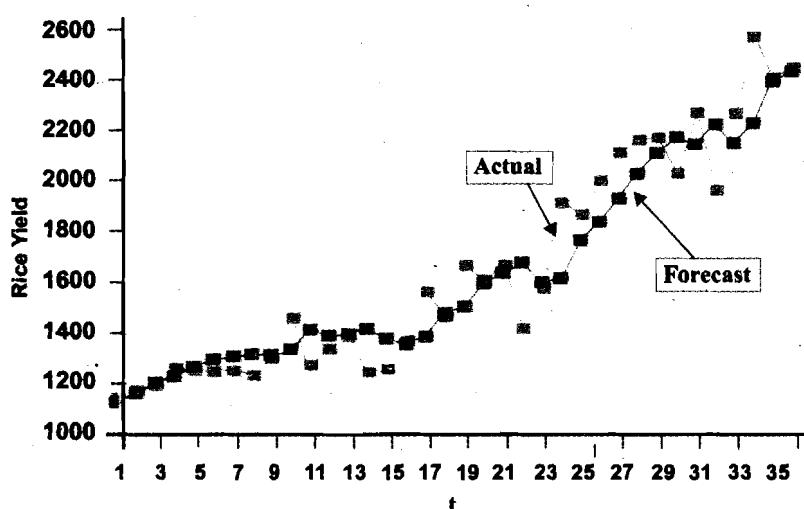


Fig. 13 : Forecast with Holt's model

Now, observe Equation (9), which gives forecasts of the immediate future. Assume that we have got the data only upto 1985, and that we wish to forecast the yields from 1986 to 1990. Note that $t = 31$ corresponds to year 1985. Equation (9) can now be generalised to make these forecasts as follows :

$$\hat{y}_{31+k} = \hat{a}_{31} + \hat{b}_{31}k, k = 1, 2, \dots \quad (10)$$

You may like to do the following exercise now.

- E23) Assume that you have the rice yield data from 1955 to 1985 only. What will your forecasts be for 1986 to 1990 if you use Holt's exponential smoothing with $\hat{a}_0 = 1090$, $\hat{b}_0 = 32$, $w_1 = 0.4$, and $w_2 = 0.01$?

With this we come to the end of our discussion on finding the trends, and hence the predicted values. Actually, to forecast correctly, we need to estimate the other components too. We will do this in one case, just to give you a flavour.

Example 4 (contd.) : In Example 4, we observed that the model suitable for forecasting is $y_t = T_t \cdot S_t \cdot I_t$. While solving E20, we found that the method of moving averages gives us the estimate $T_t = \frac{1}{3}(y_{t-1} + y_t + y_{t+1})$.

Now, to estimate the seasonal variations, we first find the estimates of $S_t I_t$ in the table below.

Table 19 : Coconut Sales Data

Period (t)	Year	Season	Coconuts Sold (y_t) (in lakhs)	Trend component (T_t)	Estimates of $S_t I_t$ (= y_t / T_t)
1	1994-95	I	14	-	-
2	1994-95	II	35	38	0.921
3	1994-95	III	65	38	1.711
4	1995-96	I	14	41	0.341
5	1995-96	II	43	45	0.956
6	1995-96	III	77	45	1.711
7	1996-97	I	16	44	0.364
8	1996-97	II	40	47	0.851
9	1996-97	III	84	47	1.787
10	1997-98	I	17	49	0.347
11	1997-98	II	46	51	0.902
12	1997-98	III	90	53	1.698
13	1998-99	I	22	54	0.407
14	1998-99	II	50	59	0.847
15	1998-99	III	105	59	1.780
16	1999-00	I	23	53	0.434
17	1999-00	II	30	58	0.517
18	1999-00	III	120	59	2.034
19	2000-01	I	27	72	0.375
20	2000-01	II	68	76	0.895
21	2000-01	III	132	-	-

We are now in a position to estimate the seasonal effects, S_I , S_{II} and S_{III} . The estimate of S_I is the average of all those estimates of $S_t I_t$ in which t corresponds to Season I. So, $S_I = \frac{S_4 I_4 + S_5 I_5 + \dots + S_{19} I_{19}}{6} = 0.378$.

We can similarly estimate S_{II} and S_{III} .

Having estimated $S_t I_t$ and S_t , we can now estimate I_t by using the equation
 $I_t = S_t I_t / S_t$.

* * *

You can wind up the example above by doing the following exercises.

-
- E24) Estimate S_{II} and S_{III} in the example above. (Note that there are seven terms in S_{II} and six terms in S_{III} .)
 - E25) Estimate I_t for all values of t in Table 9.
 - E26) Using only the first 5 years of coconut sales data (see Table 9), build the multiplicative model and estimate the trend, seasonal and irregular components.
Use this model to forecast the sales for the next two years and compare them with actual sales.
-

We shall end our discussion on forecasting here. If you would like to go into greater depth, you could refer to the books listed under 'Further Reading' in the Course Introduction. You could also see the website: www.bath.ac.uk/~masar/math0118/forecasting/node6.html.

Now let us sum up the chief points made in this unit.

10.6 SUMMARY

In this unit we have discussed the following points.

- 1) Forecasting and its importance in future planning.
- 2) Time series and its four basic components — trend, seasonal variation, cyclic variation and random variation.
- 3) The additive and multiplicative models of time series data analysis.
- 4) The method of least squares for fitting linear and non-linear trends.
- 5) Time series data analysis by the method of moving averages.
- 6) The simple exponential smoothing procedure and Holt-Winters' exponential smoothing procedure.

10.7 SOLUTIONS AND ANSWERS

- E1) The forecasting errors for Sunday to Saturday are -1, 4, -4, -9, 7, 0 and 8, respectively.
- E2) The forecasts are 43, 42, 60, 52 and 56 for Tuesday to Saturday, respectively. The forecasting errors are given in the following table.

Table 20 : Forecasting Errors

Day	Forecast	Forecasting Errors		
		First Week	Second Week	Third Week
Sun	49	-2	2	0
Mon	52	2	-3	2
Tue	43	1	-3	3
Wed	42	4	-2	-1
Thu	60	3	-1	-3
Fri	52	-1	4	-2
Sat	56	-1	-2	2

The three weeks overall profit is equal to Rs.2015/-.

- E3) The data regarding the death rate of India, the population of India, the summer temperatures in your town are examples. You can think of many other examples.
- E4)
 - i) Quarterly sales of woollen clothes, numbers of fans sold in a city month-by-month, etc.
 - ii) Number of people attending church day-wise. Here the variation is within a week, with a peak on Sunday.
- E5) The monthly consumption of rice by a family, for instance.
- E6) While doing this, remember that there are time series which may not exhibit cyclic variations. For instance, the data on death rates.
- E7) One example could be a sudden warm wave in January due to weather irregularities.
- E8)
 - i) The number of cars produced monthwise is usually large in January to March every year. Therefore, there will be a seasonal variation. Any unexpected change in the economic or labour policy can bring in irregular variations in the long-term trend. Since the demand will be more in some months and little less in others over the years, there will be a cyclic variation also. So, all the four basic components are expected to be present.
 - ii) Again, all 4 components can be present : seasonal variation due to exams, irregular variation due to the shop closing for unexpected reasons.
 - iii) Trend, cyclic and irregular components could appear. The seasonal component may not be exhibited because the time series is annual, not monthly or quarterly.
- E9) The S_t values for Mon, Wed, Thu and Fri are -2, -4, 7 and 6, respectively. So, $t_2 = -2$, $t_4 = -4$, $t_6 = 6$, $t_{16} = -2$ (Mon), $t_{26} = 7$ (Thurs).

- E10) The trend components for the weeks 6, 7 and 8 are 66, 69 and 72, respectively. To get the 6th week's forecasts, we must add 66 to the average seasonal components (given in the last column of Table 7). Similarly, we must add 69 and 72 to the seasonal components to get the other two weeks' forecasts. The forecasts are tabulated below.

Table 21 : Forecasts (Actual Sales) for 6th, 7th and 8th Weeks

Day →	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Week ↓							
6	64(59)	64(65)	59(58)	62(62)	73(79)	72(66)	70(69)
7	67(64)	67(66)	62(60)	65(57)	76(81)	75(70)	73(69)
8	70(69)	70(75)	65(67)	68(69)	79(78)	78(74)	76(73)

- E11) In 1995-96 it is 63, in 1996-97 it is 68, and so on till in 2000-01, it is 105. It is clearly increasing over the years.
- E12) Wednesdays are represented by the day numbers $t = 4, 11, 18, 25, 32, \dots$. Substituting $t = 4$ in (1), we get $T_t = 50.0$. This is the trend value of Day 4. All the required trend values by the two methods, regression and the ad hoc approach of Example 3, are tabulated below. Remember that for the ad hoc method, the starting trend value is 50.4, and thereafter it increases by 3.2 every week.

Table 22 : Trend Values By Two Methods

Method	Day (t)					
	4	11	18	25	32	...
Regression	50.0	53.4	56.8	60.2	63.6	...
Ad hoc	50.4	53.6	56.8	60.0	63.2	...

- E13) The trend equation is given by

$$T_t = -1.724 + 0.0455t, t = 61, 62, \dots \quad (11)$$

with $R^2 = 0.35$ and standard error $S_e = 0.1991$.

- E14) Two equations are fitted for you to see, and compare. One is linear regression and the other is a second degree equation. Their graphs are shown in the figure below. Here t is taken as $t = 1$ for 1901, $t = 2$ for 1911, and so on.

$$Y_t = 769 + 485.68 \times t, t = 1, 2, \dots, R^2 = 0.87, S_e = 585.41, \quad (12)$$

$$Y_t = 2305 - (282.398 \times t) + (69.825 \times t^2), t = 1, 2, \dots, R^2 = 0.99, S_e = 154. \quad (13)$$

From the values of R^2 in (12) and (13), you can see that the quadratic polynomial is a better fit than the linear one.

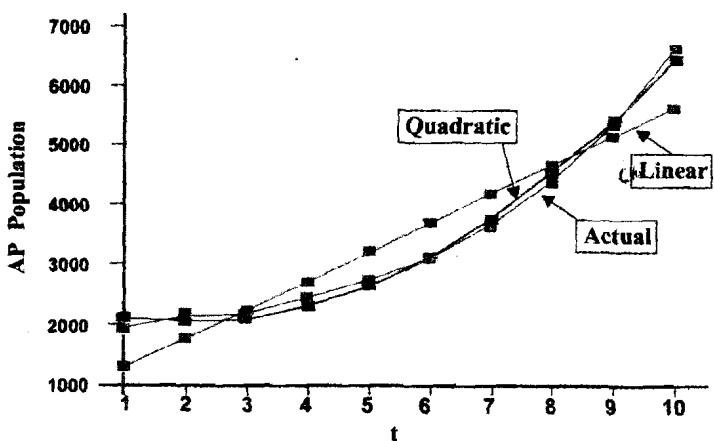


Fig. 14 : Linear and quadratic trend curves for population

- E15) a) The graph for the exponential curve fitting and the actual time series values is given below. The quadratic fit is better in this case.

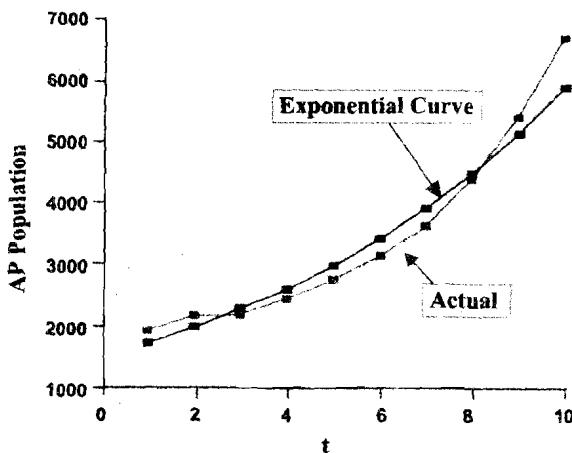


Fig. 15 : Exponential curve for AP population

- b) According to the exponential fit $T_{11} = 6,68,80,580$, and according to the quadratic fit $T_{11} = 7,64,84,470$. In this case the quadratic curve gives a better forecast.
According to the exponential fit, $T_{12} = 7,66,47,850$; and according to the quadratic fit $T_{12} = 7,95,79,040$.

- E16) Take T_t as the dependent and t as the independent variable and fit a linear regression equation. If we do this for our data, we get the estimate of T_t as $T_t = 26.1 + 2.47t$.
- E17) The first three moving averages of length 3 are 46.33, 47 and 51. These numbers should be placed against Days 2, 3 and 4, respectively.
- E18) The first and last moving averages of length 6 are 50.33 and 63.33, and they correspond to Days 3.5 and 32.5, respectively.
- E19) The moving averages are given in Columns (3), (4) and (5), respectively in Table 23 below.

Table 23

Year (t)	Crop Yield	Moving averages of length :		
		3	4	5
1955	27.23			
1956	29.27	28.27	28.7275	
1957	28.31	29.22667	29.6225	29.144
1958	30.1	29.74	29.7075	29.62
1959	30.81	30.17333	31.11	30.55
1960	29.61	31.44667	32.2725	31.838
1961	33.92	32.76	32.9625	32.532
1962	34.75	34.08	34.21	33.29
1963	33.57	34.30667	33.58	33.648
1964	34.6	33.19	33.2	33.51
1965	31.4	33.07667	33.305	33.358
1966	33.23	32.87333	31.7775	32.342
1967	33.99	31.90333	32.6	32.36
1968	28.49	32.39	33.095	33.122
1969	34.69	32.79667	32.2	32.558
1970	35.21	33.43667	32.3975	31.616
1971	30.41	31.63333	32.17	32.674
1972	29.28	31.15667	32.25	32.842
1973	33.78	32.86333	34.385	33.59
1974	35.53	36.08667	35.9775	34.638
1975	38.95	36.71	36.69	36.108
1976	35.65	37.07667	37.755	37.31
1977	36.63	37.35667	36.69	37.142
1978	39.79	37.03667	36.7775	36.552
1979	34.69	36.82667	37.18	37.07
1980	36	36.31	36.3275	37.02
1981	38.24	36.87333	38.0625	37.388
1982	36.38	38.75	37.8075	37.446
1983	41.63	37.66333	36.8775	37.15
1984	34.98	37.04333	36.43	36.42
1985	34.52	34.69667	34.04	35.558
1986	34.59	33.72667	35.84	35.668
1987	32.07	36.28	37.725	37.084
1988	42.18	38.77	39.1675	38.252
1989	42.06	41.53333	40.99	39.206
1990	40.36	40.59333	39.455	40
1991	39.36	38.58667	37.8075	38.658
1992	36.04	36.95667	36.81	37.52
1993	35.47	35.96	36.2	36.832
1994	36.37	36.25333		
1995	36.92			

Let us consider the curves given by these in Fig. 16 below.

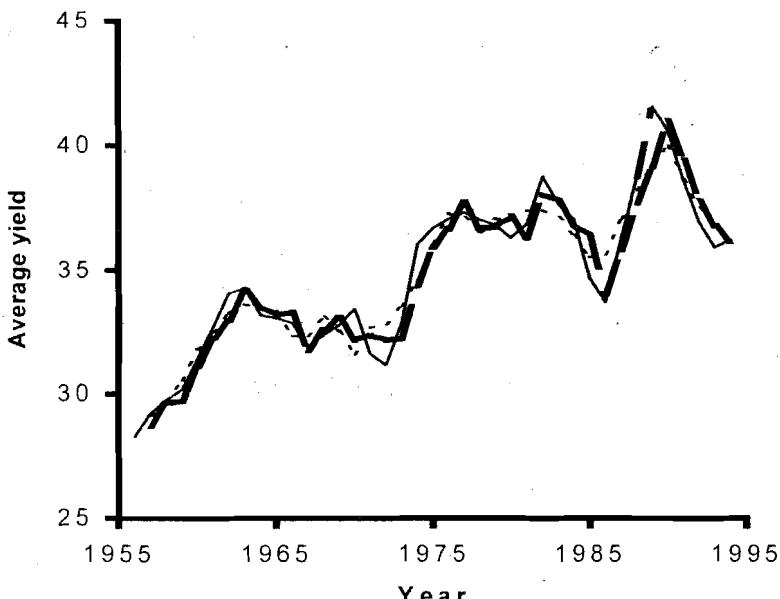


Fig.16

Looking at the curves, we see that the cycles are usually of 5-year periods because the moving averages of length 5 do not show the cyclical variation. The S-component, in any case, can be excluded because the data is annual and doesn't allow us to observe seasonal variation.

- E20) Consider the average of y_1, y_2 and y_3 . This is the average of sales belonging to three different seasons. Similarly, the average of y_2, y_3 and y_4 is also the average of sales belonging to three different seasons. This way, we can compute the average of y_{t-1}, y_t and y_{t+1} for $t = 2, 3, \dots, 20$ (rounded off to the nearest integer). Since these averages have the effect of all seasons and the random effect, they are treated as the estimates of T_t . We have obtained the values in the table below.

Table 24 : Coconut Sales Data

Period (t)	Year	Season	Coconuts Sold (y_t) (in lakhs)	Trend component (T_t)
1	1994-95	I	14	-
2	1994-95	II	35	38
3	1994-95	III	65	38
4	1995-96	I	14	41
5	1995-96	II	43	45
6	1995-96	III	77	45
7	1996-97	I	16	44
8	1996-97	II	40	47
9	1996-97	III	84	47
10	1997-98	I	17	49
11	1997-98	II	46	51
12	1997-98	III	90	53
13	1998-99	I	22	54
14	1998-99	II	50	59
15	1998-99	III	105	59
16	1999-00	I	23	53
17	1999-00	II	30	58
18	1999-00	III	120	59
19	2000-01	I	27	72
20	2000-01	II	68	76
21	2000-01	III	132	-

E21) $\hat{a}_3 = 0.02y_3 + 0.98\hat{a}_2 = 0.02 \times 847 + 0.98 \times 946.14 = 944.16.$

The corresponding forecast error,

$$\hat{e}_4 = y_4 - \hat{a}_3 = 1063 - 944.16 = 118.84.$$

$$\hat{a}_4 = 0.02y_4 + 0.98\hat{a}_3 = (0.02 \times 1063) + (0.98 \times 944.16) = 946.54.$$

The corresponding forecast error is

$$\hat{e}_5 = y_5 - \hat{a}_4 = 1030 - 946.54 = 83.46.$$

E22)

Table 25 : Exponential Smoothing of Banks Data

Year	No. of Branches	Forecast	Error
81	5	4.00	1.00
82	3	4.01	-1.01
83	3	3.999	-0.99
84	4	3.99	0.01
85	3	3.99	-0.99
86	6	3.98	2.02
87	4	4.00	0.00

E23)

Table 26 : Forecasts of Rice Yield For 1985 to 1990

Year	t	Rice Yield	Forecast	Error
1986	32	1951	2114	-163
1987	33	2258	2146	112
1988	34	2572	2178	394
1989	35	2403	2210	193
1990	36	2442	2242	200

E24)

$$S_{II} = \frac{S_2 I_2 + \dots + S_{20} I_{20}}{7} = 0.841$$

$$S_{III} = 1.787$$

E25)

Table 27

Period (t)	Year	Season	Coconuts Sold (y_t) (in lakhs)	Trend component (T_t)	Estimates of $S_t I_t$ (= y_t / T_t)	Estimates of S_t	Estimates of $I_t (= S_t I_t / S_t)$
1	1994-95	I	14	-	-	0.378	-
2	1994-95	II	35	38	0.921	0.841	1.095
3	1994-95	III	65	38	1.711	1.787	0.957
4	1995-96	I	14	41	0.341	0.378	0.903
5	1995-96	II	43	45	0.956	0.841	1.136
6	1995-96	III	77	45	1.711	1.787	0.958
7	1996-97	I	16	44	0.364	0.378	0.962
8	1996-97	II	40	47	0.851	0.841	1.012
9	1996-97	III	84	47	1.787	1.787	1.0
10	1997-98	I	17	49	0.347	0.378	0.918
11	1997-98	II	46	51	0.902	0.841	1.072
12	1997-98	III	90	53	1.698	1.787	0.95
13	1998-99	I	22	54	0.407	0.378	1.078
14	1998-99	II	50	59	0.847	0.841	1.007
15	1998-99	III	105	59	1.780	1.787	0.996
16	1999-00	I	23	53	0.434	0.378	1.148
17	1999-00	II	30	58	0.517	0.841	0.615
18	1999-00	III	120	59	2.034	1.781	1.138
19	2000-01	I	27	72	0.375	0.378	0.992
20	2000-01	II	68	76	0.895	0.841	1.064
21	2000-01	III	132	-	-	1.781	-

- E26) Let us form a table, based on the first 15 readings of Table 9. Since we need to forecast the sales, we should use regression for finding the trend. (Remember, MA is not a useful method for forecasting the trend).

To estimate S_t , we use the same procedure used in E25. Let us use the ad hoc approach (see Example 3) for determining the long-term trend component.

Now, the average sales for the first year are $\frac{14 + 35 + 65}{3} = \frac{114}{3}$, for the

2nd year are $\frac{134}{3}$, $\frac{140}{3}$ for the 3rd year, $\frac{153}{3}$ for the 4th year, and $\frac{177}{3}$ for the 5th year. So, the average increase in trend is

$\frac{1}{4} \left(\frac{20}{3} + \frac{6}{3} + \frac{13}{3} + \frac{24}{3} \right) = \frac{21}{4} = 5.25$ units. Therefore, the 5th, 6th, and 7th columns in the table will be as shown.

Table 28

Period (t)	Year	Season	Coconuts Sold (y_t) (in lakhs)	Estimates of T_t	Estimates of S_t	Estimates of I_t ($=y_t - S_t - T_t$)
1	1994-95	I	14	38	0.358	1.029
2	1994-95	II	35	38	0.895	1.029
3	1994-95	III	65	38	1.727	0.990
4	1995-96	I	14	43.25	0.358	0.904
5	1995-96	II	43	43.25	0.895	1.111
6	1995-96	III	77	43.25	1.727	1.031
7	1996-97	I	16	48.5	0.358	0.921
8	1996-97	II	40	48.5	0.895	0.921
9	1996-97	III	84	48.5	1.727	1.003
10	1997-98	I	17	53.75	0.358	0.883
11	1997-98	II	46	53.75	0.895	0.956
12	1997-98	III	90	53.75	1.727	0.97
13	1998-99	I	22	59	0.358	1.042
14	1998-99	II	50	59	0.895	0.947
15	1998-99	III	105	59	1.727	1.030

Using these estimates, the forecast for the three seasons in the next two years, ignoring the irregular component, are $(64.25)(0.358)$, $(64.25)(0.895)$, ..., $(69.5)(1.727)$, i.e., 23, 58, 111, 25, 62, 120 (rounded off to the nearest integer). The forecasting errors are 0, -28, 9, 2, 6, 12, respectively.

UNIT 11 STATISTICAL QUALITY CONTROL

Structure	Page No.
11.1 Introduction Objectives	77
11.2 Concept of Quality Nature of Quality Control	78
11.3 Statistical Process Control Concept of Variation Control Charts Control Charts For Variables Process Capability Analysis Control Charts For Attributes	79
11.4 Acceptance Sampling Sampling Plan Concepts Single Sampling Plans	91
11.5 Summary	98
11.6 Solutions/Answers	99

11.1 INTRODUCTION

A widely accepted definition of the *quality* of a product is **its fitness for use for its intended purpose**. For example, a ball pen should write well throughout its life.

Besides, the cap should not be loose, neither it should leak nor break easily, etc., are some of the other features. Again, for a cricket ball, some of the quality characteristics are like its *weight, size, shining, quality of stitches*, etc. And, for a water tap washer, these are its *thickness, inner diameter, outer diameter*, etc.

The *quality* of a product is assessed by the totality of its features. Have you ever thought of when and where products are made? We only curse the products (and the people who made them) whenever these products give us trouble or displeasure. Rarely we think how these products are made or what precautionary measures have been taken during the production process to ensure that the products are of *good quality*.

Who is responsible for the quality of a product? Obviously, it is the manufacturer of the product. In this unit, we shall discuss some simple statistical tools, which are extensively used in the production process to ensure the quality of products. The most effective use of *Statistical Quality Control (SQC)*, in short) generally requires cooperation among those responsible for the three different types of functions: *specification, production, and inspection*.

In Sec.11.2, you will get introduced to the concept of quality and learn about methods used in the process of controlling and systematically improving quality of a product. In Sec.11.3, we shall discuss the primary tools of *statistical process control* that are useful in monitoring the quality aspects of a product. In Sec.11.4, you will learn about *acceptance sampling plans* - a technique used in ensuring that the produced products conform to the specified quality standards.

Objectives

After reading this unit, you should be able to

- explain the term *quality* and the phrase *statistical quality control*;
- describe the concept of *variation, chance and assignable causes*;

Quality Control (QC, in short) consists of procedures and methodologies which ensure that the quality characteristics of a product conform to its specifications.

- construct and interpret *control charts* for variables and attributes;
- estimate *process capability* from control charts data;
- describe *acceptance sampling plans*;
- interpret and use *OC curves* in determining *acceptable quality level*, *producer's risk*, *consumer's risk*, and *lot tolerance percentage defective*;
- describe *single sampling plan* and construct some simple single sampling plans with help of a *binomial nomograph*.

11.2 CONCEPT OF QUALITY

Everyday, from the time we get up and till we go back to bed, each one of us is busy with a number of activities. And, we depend on number of objects to carry out these activities. For example, we need a tooth brush, a tooth paste, wash basin or a water tap, the soaps and detergents, the stoves that we use for cooking, the vehicles that we use for going to our offices, electrical bulbs, phones, medicines and what not?

Quality of any product depends upon certain **characteristics** related to the product and the materials that go into it. For example, in the case of a ball pen, if the diameter of the ball is undersized then the pen is bound to leak. Similarly, if the refill length is oversized, then the press button may not work properly. So, *ball diameter* and *refill length* are two **quality characteristics** for a simple product like ball pen.

In fact, a ball pen has many more quality characteristics. But, unlike diameter and length, not all quality characteristics are *measurable*. Recall, *non-measurable characteristics* are called **attributes** and the *measurable* ones are called the **variables**.

11.2.1 Nature of Quality Control

Let us continue our discussion with the example of a ball pen. As you may agree, certain brands of ball pens stop writing from the second or third day onwards; some write well only on certain types of papers; some write well on almost any kind of paper and write till the last drop of ink remains in the refill. When a pen writes well, we say it is of *good quality*.

How we decide the *quality* of a product? At the time when a product is designed, certain *specifications* or *levels of tolerances* are established on all important quality characteristics of the product. For example, in case of a ball pen, the specifications on refill length may be that it should lie between 9.80 cms and 10.20 cms. So, if we can ensure that all the quality characteristics of a product are maintained within their specified limits, then automatically the quality of that product will be good.

But, we know that a *manufacturing process* is an interaction among people, equipments, materials, methods and environment, wherein output could be either another product or a component that goes into an end product. Thus, the performance of a *process* is indicated by the quality of its output and can be assessed by examining the *quality characteristics* of the output. So, a *process* is operated in such a way that the quality characteristics of output product is maintained at desired levels. This is called **controlling a process**.

In 1924, **Walter A. Shewhart** of Bell Telephone Laboratories introduced *statistical control charts* as a tool for controlling quality of industry products. From then onwards, people slowly started recognising the use of statistical techniques in quality control. Today, it is known world over that statistical techniques are not only indispensable for quality control but also play a very crucial role in all other facets of quality related matters.

Cooking in hotels is an example of a manufacturing process.

Statistical Quality Control techniques can be broadly divided into two categories: (i) *Statistical Process Control (SPC)*, in short) techniques; and (ii) *Acceptance Sampling*.

SPC techniques are widely used in almost any manufacturing process and are very useful in solving real situation problems, achieving *process stability*, and *making continuous improvements* in product quality. The most important among these are **control charts**. We shall discuss *control charts* in the next section.

In many situations, however, one or more components of a product are bought from outside agents and the manufacturer does not have a direct control over the quality of the components. Then, in such cases, **acceptance sampling techniques** are useful in ensuring that the bought out components conform to specified quality levels.

In the next section, we would discuss *SPC techniques* and *acceptance sampling* will be discussed in Sec.11.4.

SPC techniques have wide applications in non-manufacturing processes as well.

11.3 STATISTICAL PROCESS CONTROL

Statistical Process Control is a methodology used for understanding and monitoring a process by collecting the data on quality characteristics periodically from the process, analysing them and taking necessary actions based on the analysis results.

From now onwards, we will focus our discussion by referring to M/s BP & Company, a ball pen manufacturing company.

One of the sections in this company produces refills for the ball pens. The specifications on the refill length, one of the quality characteristics of refill, are 10 ± 0.2 cms. While producing the refills how does one ensure that their lengths conform to these specifications?

Refill length problem

11.3.1 Concept of Variation

As you may agree, however well the process is maintained, certain amount of variation in the lengths of the refills is unavoidable. But, if this process is operated under stable conditions i.e., *machine settings are same, quality of the materials used is same, operators are equally experienced*, etc., then the quality characteristics such as refill length normally exhibit a specific *patterns of variation*.

Pattern of variation means *statistical distribution*.

That is, if a process is operating under stable conditions then the amount of variation in the quality characteristic is usually small and is a result of several small causes. These small causes are known as *chance causes* and are usually inevitable.

Chance Causes

The resulting variation is called the *chance cause variation* (or the *inherent variation*) of the process. In practice, we find that most processes are often disturbed inadvertently or otherwise.

On the other hand, a change in the machine settings, sudden drop in the quality of raw material or induction of a new operator due to an absence of the regular operator, etc., are some of the causes that might disturb a *stable process*. The reasons for variation outside this stable process may be discovered and corrected. These causes are known as *assignable causes*.

When the *assignable causes* are prevailing in a process, the process becomes unstable and this is reflected in the behaviour of the quality characteristic. That is, there will be frequent changes in the distribution of the quality characteristic. As a result, there will be more variation in the data. This variation is called the *variation due to assignable causes*.

Assignable Causes

It is important to remember that when a process is operating only under *chance causes*,

we say that the process is statistically stable or that *the process is under statistical control.*

The power of *Shewhart control charts* lies in its ability to separate out assignable causes of quality variation. So, a control chart is used to monitor the stability of a process and alert us as and when an assignable cause creeps into the process. Also, control charts are very useful in detecting gradual improvement or deterioration in a process.

11.3.2 Control Charts

On-line process control means monitoring a process by periodically examining sample outputs of the process and taking corrective actions as and when necessary.

The concept of *control charts* is one of the most powerful techniques for *on-line process control*. Besides monitoring, control charts are useful also in the evaluation of capability of a process and in making continuous improvements in the process.

Recall the *refill length problem* mentioned above. We can use control chart technique to effectively solve this problem. But, firstly, let us see how a control chart is constructed.

Typically, a control chart is a two-dimensional graph in which *x-axis represents the sample numbers* and *y-axis represents a quality characteristic*. It has a solid *center line (CL)* and two dotted lines called *upper control limit (UCL)* and *lower control limit (LCL)* (see Fig. 1).

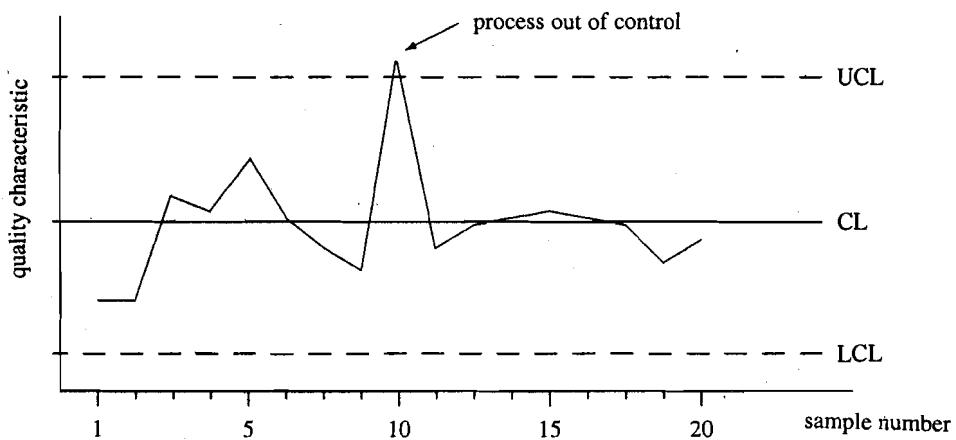


Fig. 1

Thus, the **construction of control charts** involves *collecting samples periodically from the process, computing the quality characteristic for each sample and plotting it against the sample number. The consecutive points are joined by line segments.*

As long as the plotted points are within the *upper* and *lower control limits* and do not exhibit any specific patterns, we have no evidence that the process is *not under statistical control*. When a point falls outside the control limits (below LCL or above UCL), it is a cause and indicates the presence of an assignable cause with a high probability.

However, *control chart* cannot tell us what went wrong with a process when something has gone wrong. It will only indicate that possibly something has gone wrong with the process. In fact, it is the responsibility of supervisor or QC manager to find out what has gone wrong.

In most situations, a *quality characteristic follows a normal distribution* or can be approximated by a normal distribution. Also, we know that the probability of a normally distributed random variable taking values below $\mu - 3\sigma$ or above $\mu + 3\sigma$, where μ is the mean and σ is the standard deviation, is very low (equal to 0.0027).

Therefore, if an observation falls outside *3σ limits*, it is logical to suspect that possibly something might have gone wrong. For this reason, *the control limits on a control chart are set up using 3σ limits*. Consequently, when a point falls outside the control limits on a control chart, it is more likely that it is due to the presence of an assignable cause,

Depending on the nature of quality characteristics, control charts are divided into two categories: (i) *control charts for variables*; and (ii) *control charts for attributes*.

Before we proceed further with our discussion, try the following exercise.

-
- E1) For what type of quality characteristic do you think is a control charts for variables suitable? Cite some quality characteristics that need control charts for attributes.
-

Control charts for variables are adopted in situations where the quality characteristic is of measurable type. In the next part of the section, we shall discuss with you \bar{x} - R charts – a type of control charts for variables.

11.3.3 Control Charts For Variables

Once again we shall refer to the refill length problem to explain these charts. Consider the data on refill length as given in Table 1. Observe that the samples are taken at 30-minute intervals. Here, each sample corresponds to lengths of *five refills* produced at the time of collecting the sample. Each sample is called a **subgroup**.

Table 1 : Refill length data

S.No.	Date	Time	Subgroup					Average	Range
			1	2	3	4	5		
1	23.12.97	08:00	10.11	10.08	10.14	10.10	10.15	10.116	0.07
2		08:30	10.08	10.08	10.12	10.13	10.06	10.094	0.07
3		09:00	10.07	10.22	10.01	10.11	10.07	10.096	0.21
4		09:30	10.21	10.10	10.09	10.13	10.02	10.110	0.19
5		10:00	10.12	9.98	9.91	10.05	10.17	10.046	0.26
6		10:30	10.17	10.14	10.08	10.06	10.23	10.136	0.17
7		11:00	10.10	10.11	10.21	10.05	10.22	10.138	0.17
8		11:30	10.10	10.06	10.23	10.14	9.97	10.100	0.26
9		12:00	10.10	9.96	10.13	10.14	10.04	10.074	0.18
10		12:30	10.05	10.19	10.13	10.10	10.08	10.110	0.14
11	24.12.97	08:00	10.08	10.05	10.05	10.08	10.16	10.084	0.11
12		08:30	9.91	10.21	10.00	10.02	10.29	10.086	0.38
13		09:00	10.11	9.98	9.97	10.04	10.08	10.036	0.14
14		09:30	10.08	10.21	10.13	10.16	10.04	10.124	0.17
15		10:00	9.99	10.14	9.96	10.09	10.07	10.050	0.18
16		10:30	10.17	10.18	10.04	9.99	10.11	10.098	0.19
17		11:00	10.06	9.92	10.10	10.06	10.02	10.032	0.18
18		11:30	10.16	10.12	10.16	10.02	10.19	10.130	0.17
19		12:00	10.14	10.04	10.14	10.02	10.07	10.082	0.12
20		12:30	10.08	10.07	9.97	10.09	10.12	10.066	0.15

Thus, in above table, we have data for 20 *subgroups*. The variation in the five observations of any subgroup can be attributed only to *chance causes* because the *five observations* correspond to *five refills* that were produced almost at the same time and it is very unlikely that they were affected by any assignable causes in such a short span of time.

Next, let us talk about the *frequency of sampling*. Supposing the samples are taken once in every five minutes instead of every 30 minutes, then we are not going to find big differences or changes in consecutive subgroups. So, *too frequent sampling is an unnecessary labour*.

While keeping these points in mind, one has to decide the *selection of subgroups* and *their frequency* in such a way that the variation in observations within a subgroup is only due to chance causes and the variation among subgroups is likely to be affected by assignable causes. The subgroups selected in this way are called the **rational subgroups**.

Here, **LSL** and **USL** stands for the *lower and upper specification limits*, respectively.

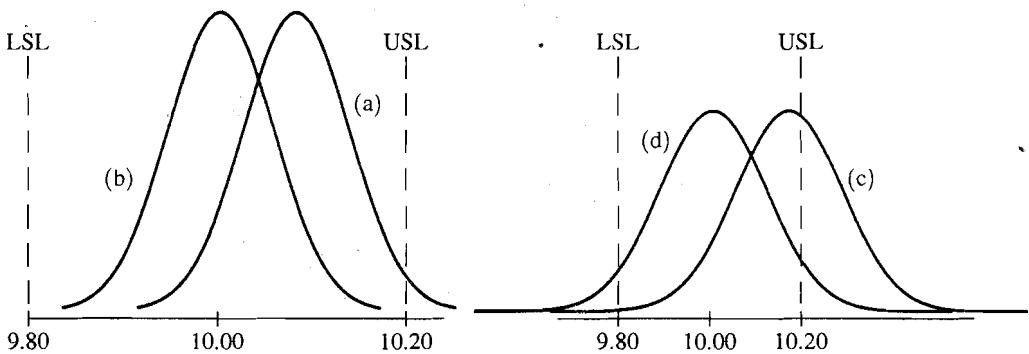


Fig. 2

The four curves (a – d) in Fig. 2 describe four different distributions for refill lengths. Observe that, for the curve (a), the mean length is high even though the variation is low. So, if our process produces refills like this, then some percentage of the refills do not conform to the length specifications. However, if we can adjust the process so that the mean length is equal to 10 cms, then our process will be good. Then, in this situation, the refill length distribution will look like the one in (b).

On the other hand, if the distribution is given by (c) then even after adjusting the mean to the target there will be *non-conformance* with regard to length (see curve (d)). So, in this case, we must improve the process to reduce the variation.

The moral of the story is that, *when the quality characteristic is a variable, it is essential to control both mean and variation*. For this reason, *two separate control charts* – one for controlling the mean and the other for controlling variability, are maintained for a measurable quality characteristic.

In quality control terminology, when a process is stable, the mean of a quality characteristic under study is referred to as **process mean** and its variability is referred to as **process variability**.

Here, we denote the process mean by μ and process standard deviation by σ .

Now, let x_1, x_2, \dots, x_5 be five independent observations of a *subgroup* on the refill length. If the subgroup is *rational*, we can expect that x_1, x_2, \dots, x_5 are random variables having same mean and standard deviation.

Let \bar{x} and R denote the *average* and *range*, respectively, of these five observations. Then, \bar{x} is called the **subgroup average** and R the **subgroup range**.

Two separate charts are maintained in an \bar{x} - R chart: (i) \bar{x} - chart; and (ii) R - chart. In the \bar{x} - chart, used for controlling the *process mean*, we plot the *sample averages against the sample numbers*; and, in the R - chart, used for controlling *process variability*, we plot the *sample range against the sample number*.

Recall, the *control limits* are worked out based on the 3σ -limits concept. Also, from what you read in Unit 4, we know if x_1, x_2, \dots, x_n are independent random variables with mean μ and standard deviation σ , then \bar{x} has mean μ and its standard deviation is equal to $\frac{\sigma}{\sqrt{n}}$ (by central limit theorem).

As such, since we plot the sample averages on the \bar{x} - chart, we should construct the *center line* and *control limits* using the mean and standard deviation of the averages and not of the individual observations. Thus, the CL, LCL and UCL for the \bar{x} - chart are

given by

$$CL = \mu, LCL = \mu - 3 \frac{\sigma}{\sqrt{n}} \text{ and } UCL = \mu + 3 \frac{\sigma}{\sqrt{n}}, \quad (1)$$

where n is the sample size (For example, $n = 5$ for the refill length problem). And, the estimates of μ and σ are given by $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = \bar{R}/d_2$, where \bar{x} is the average of all subgroup averages, \bar{R} is the average of subgroup ranges, and d_2 is a constant depending on n . Substituting these estimates in Eqn.(1), we get

$$CL = \bar{x}, LCL = \bar{x} - 3 \frac{\bar{R}}{d_2 \sqrt{n}} = \bar{x} - A_2 \bar{R}, UCL = \mu + 3 \frac{\bar{R}}{d_2 \sqrt{n}} = \bar{x} + A_2 \bar{R}, \quad (2)$$

$$\text{where } A_2 = \frac{3}{d_2 \sqrt{n}}.$$

Similarly, the control limits for an R - chart are also worked out based on the 3σ limits concept and these are given by

$$CL = \bar{R}, LCL = d_3 \bar{R} \text{ and } UCL = d_4 \bar{R}, \quad (3)$$

where d_3 and d_4 are constants depending n . In our subsequent discussion, we shall make use of the values of d_2 , d_3 , d_4 and A_2 as given in Table 2.

Table 2 : Control chart constants

Sample size, n	d_2	A_2	d_3	d_4
2	1.128	1.88	0	3.27
3	1.693	1.02	0	2.57
4	2.059	0.73	0	2.28
5	2.326	0.58	0	2.11
6	2.534	0.48	0	2.00
7	2.704	0.42	0.08	1.92

For example, when $n = 5$, $d_2 = 2.326$, $A_2 = 0.58$, $d_3 = 0$ and $d_4 = 2.11$. Let us use these values to solve the following problem.

Problem 1. Estimate μ and σ for refill length data and work out the control limits for \bar{x} and R - charts.

Solution. Using data given in Table 1, we get

$$\bar{x} = \frac{10.116 + 10.094 + \dots + 10.066}{20} = \frac{201.808}{20} = 10.09, \text{ and}$$

$$\bar{R} = \frac{0.07 + 0.07 + \dots + 0.15}{20} = \frac{3.51}{20} = 0.175.$$

Then, the estimates for μ and σ are given by

$$\hat{\mu} = 10.09 \text{ and } \hat{\sigma} = \frac{0.175}{2.326} = 0.0752.$$

Therefore, control limits for \bar{x} - chart are given by

$$CL = \bar{x} = 10.09; LCL = \bar{x} - A_2 \bar{R} = 10.09 - 0.58 \times 0.175 = 9.99; \text{ and}$$

$$UCL = \bar{x} + A_2 \bar{R} = 10.09 + 0.58 \times 0.175 = 10.19.$$

Similarly, the control limits for R - chart are given by

$$CL = \bar{R} = 0.175; LCL = d_3 \bar{R} = 0.0 \times 0.175 = 0.0; \text{ and}$$

$$UCL = d_4 \bar{R} = 2.11 \times 0.175 = 0.369.$$

Now, you try the following exercise.

- E2) A cricket ball manufacturing company wants to maintain control charts for the weight of the balls. Twenty-five samples, each of size 4, were collected. The *sum of sample averages* and the *sum of sample ranges* were found to be 7575 grams and 154 grams, respectively. Estimate the process mean and standard deviation, and compute the control limits for \bar{x} and R - charts.

At this stage, we may call these control limits as **trial control limits** because we do not know whether the data we have analysed correspond to a *stable process* or not.

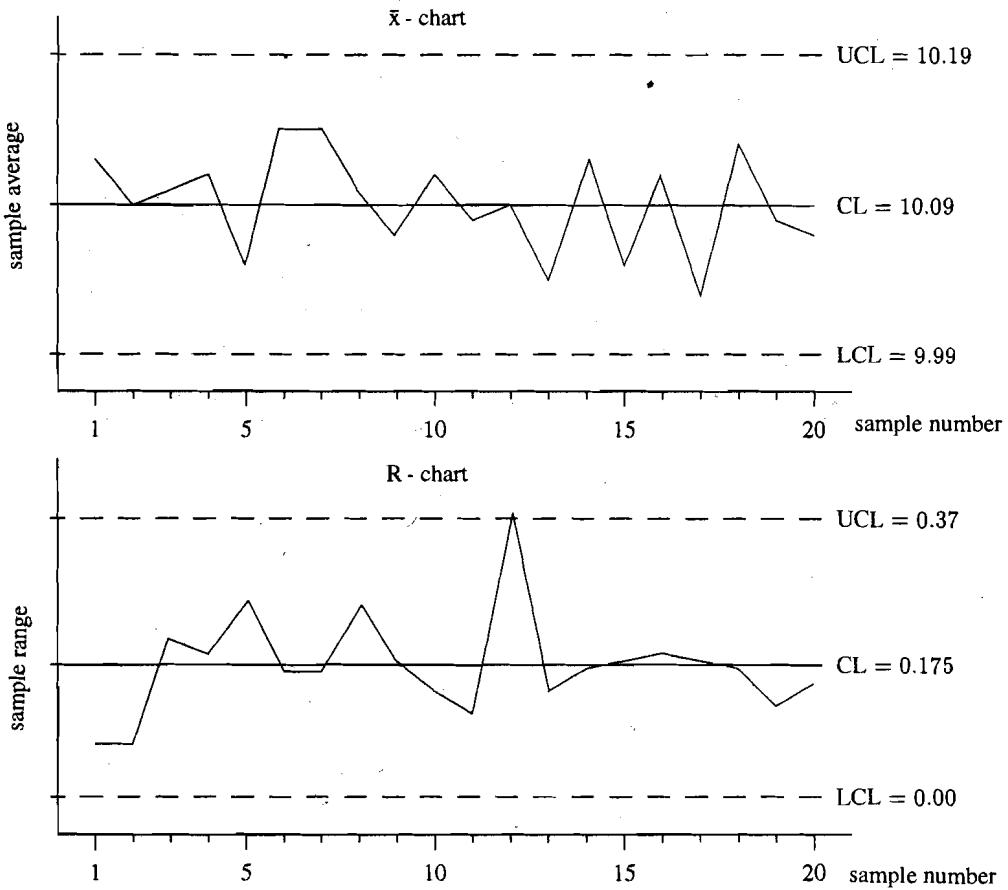


Fig. 3

Note that the 12th point on the R - chart indicates an out of control situation with regard to variability. The 12th subgroup might not be similar to the other subgroups, indicating the process might not be stable.

Though the actual cause could not be ascertained, there must have been one or more *assignable causes* which have caused this high variation. In this situation, it is better to recalculate the control limits omitting the susceptible subgroup number 12.

You try to do that in the following exercise.

-
- E3) Discard the data of 12th subgroup in Table 1 and estimate μ and σ . Compute the control limits for \bar{X} and R - charts with the remaining 19 subgroups. Do the data indicate statistical control without 12th subgroup?
-

While *homogenizing*, if we have to discard more than 25% of the subgroups, then it is better to discard the entire data and collect fresh data. But, while collecting data for control chart analysis, it is important to ensure that *process conditions remain the same throughout the data collection period*.

Now, from the solution of E3, we know that if we redraw the control charts after eliminating the 12th subgroup then both \bar{X} and R- charts exhibit a state of statistical control. But then, does it mean that the process meets the requirements?

Problem 2. Do the refills produced by this process conform to their length specifications?

Solution. Let X stands for the length and we assume that it is distributed normally with the mean and standard deviation as 10.09 and 0.075, respectively. Then, we find that

$$P[X > 10.20] = P \left[\frac{X - 10.09}{0.075} > \frac{10.20 - 10.09}{0.075} \right] = P[Z > 1.47] = 0.07,$$

where Z is the *standard normal distribution*. This means about 7% of the refills produced by the process are oversize. It is clear from the \bar{X} - chart that the process mean

After observing this, the QC manager suspected that there was a setting problem in the refill cutting machine. But, before carrying out any investigation, he decided to analyse the variation aspect too.

11.3.4 Process Capability Analysis

Supposing that QC manager has corrected the process so that the mean is as desired, do you think that the refills will conform to their length specifications? The answer is **NO**.

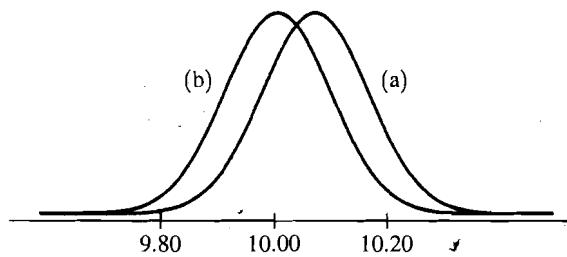


Fig. 4

In Fig. 4, curve (a) represents the process before adjusting for the mean and (b) represents the process after adjusting the mean. Note that there is no change in the variability. Thus, we will continue to have rejections even after adjusting for mean unless the process variability itself is reduced.

At this stage, we have to answer two questions: (1) What is the existing variability? (2) How much should we reduce it by? To answer these questions, we need the following definitions.

Definition. *Total tolerance* of a measurable quality characteristic, denoted by T , is given by the difference $T = USL - LSL$, where *USL* and *LSL* are the *upper* and *lower specification limits*, respectively.

For example, as $LSL = 9.80$ cms and $USL = 10.20$ cms for refill length problem, so, the *total tolerance* $T = 0.4$ cms in this case.

Definition. When a process is under statistical control, its *process capability* is given by 6σ , where σ is the *process standard deviation*.

The first question that we raised above can be answered by specifying an estimate of the process capability. Here, in our situation,

$$\text{estimate of the process capability} = 6\bar{R}/d_2.$$

For refill length problem, the specification limits are 10 ± 0.2 cms. So, to avoid rejections, we must necessarily have a process for which the 3σ limits lie within the specification limits after setting the mean at the target (see Fig. 5).

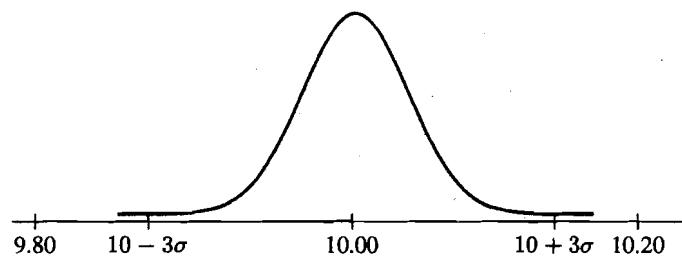


Fig. 5

The concept of 3σ -limits is used to define the **process capability**.

In other words, the *process capability must be less than the total tolerance*. Therefore, the answer to the second question we raised above is that the *process capability should not exceed the total tolerance*.

Definition. The *process capability ratio* of a stable process is the ratio of total tolerance to the process capability and is denoted by C_p . That is,

$$C_p = \frac{\text{total tolerance}}{\text{process capability}} = \frac{USL - LSL}{6\sigma}$$

Thus, by what we have said above, if $C_p < 1$, then the process is bound to produce rejections even when the mean is set on target. And, if $C_p \geq 1$, the rejection percentage will be almost zero, provided the mean is at the target.

An estimate of C_p can be obtained by substituting the estimate for σ in the above formula. For example, an estimate of C_p for the refill length problem under study is given by

$$\hat{C}_p = \frac{10.20 - 9.80}{6 \times 0.075} = 0.89.$$

Since the estimate of C_p is less than 1, we may infer that the *refill length process is not capable*. More generally, even if we get the estimate of C_p slightly more than 1, we may still consider the process incapable. This is because our estimate for C_p may be an underestimate due to sampling fluctuations.

Try the following exercise now.

E4) Give an example of

- a) a process whose $C_p = 1.5$ but has high rejections on USL side; and
- b) a process whose $C_p = 1.5$ but has high rejections on LSL side.

(Hint: Specify the process parameter μ and σ .)

With above analysis in hand, the QC manager carried out an investigation of the process and found out that a unit of the machine used for setting the refill length was not properly calibrated. He also found that certain parts in the machine were worn out which were causing vibrations in the length cutting machine.

Subsequently, he got the unit recalibrated and replaced the worn out parts and collected five samples from the process. The new data is given in Table 3.

Table 3 : Refill length data after corrective action.

S.No.	Date	Time	Subgroup					Average	Range
			1	2	3	4	5		
21	6.1.98	08:00	9.88	10.02	9.94	9.86	10.04	9.925	0.18
22		08:30	9.99	10.08	10.03	10.01	10.04	10.028	0.09
23		09:00	10.00	10.06	9.98	10.03	10.01	10.018	0.08
24		09:30	9.94	9.92	9.95	10.00	10.02	9.953	0.10
25		10:00	10.03	10.05	10.08	10.08	10.09	10.060	0.06

The new points are plotted on the charts shown in Fig. 1 and new charts now look like as in Fig. 6.

Note the clear distinction between the behaviours of the first 20 points and the last 5 points on both \bar{x} and R - charts. The difference is the effect of a change in the process after the first 20 points. And the change is the result of the corrective actions taken by the manager.

Since some changes were made in the process, we must reconstruct the control charts with new data. To complete the task, manager has collected 15 more samples. The sample averages and ranges are summarised in Table 4.

Table 4 : Summary of 15 more samples on refill length data

S.No.	Average	Range	S.No.	Average	Range	S.No.	Average	Range
26	10.035	0.14	31	10.053	0.08	36	10.048	0.18
27	9.995	0.14	32	9.973	0.08	37	10.030	0.14
28	10.020	0.08	33	9.983	0.20	38	10.048	0.22
29	9.970	0.13	34	10.020	0.16	39	10.073	0.15
30	9.970	0.10	35	9.990	0.11	40	9.985	0.23
Total	49.99	0.59	Total	50.019	0.63	Total	50.184	0.92

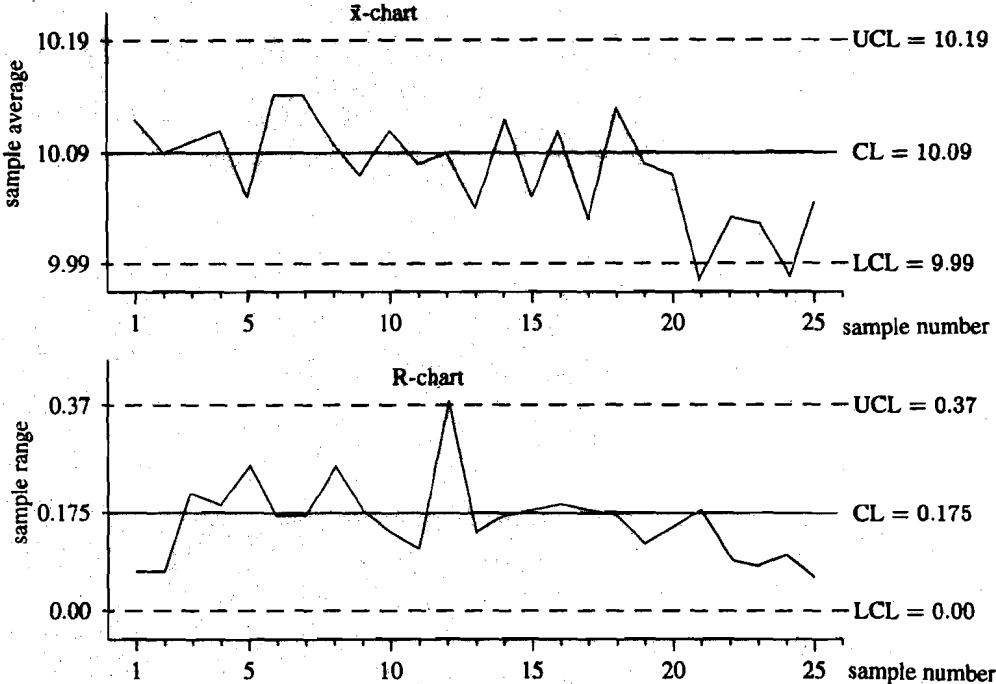


Fig. 6

You can help and advise the manager with the above data, provided you do the following exercise.

- E5) Combining the information provided in Table 3 and Table 4, do the following.
- Estimate the new process mean and standard deviation;
 - Construct the control charts;
 - State whether the process is under statistical control;
 - Estimate the process capability and process capability ratio; and
 - State your advice to the manager.

Next, let us talk about some control charts used for *non-measurable* quality characteristics.

11.3.5 Control Charts for Attributes

When products are inspected, they are classified into good and defective products. A defective product is one that has one or more defects. The performance of a process is often assessed by the *proportion of defective items* produced by the process or by counting the *number of defects per unit* of the product.

Control charts used in these situations are known as **attribute control charts**. Here, we discuss the following four such type of charts.

- p and np charts for the control of defective products;
- c charts for the control of number of defects per unit.

In case of BP & Co., 100 refills are selected at random from the process each day and are inspected for all quality characteristics. Based on the inspection results, each refill is classified as *good* or *defective*.

A defect is a *nonconformity* with respect to any of the quality characteristics of the product.

p-charts

So, each day's sample of 100 refills is taken as a subgroup. The results of 14 days sample collection are given in Table 5. Here, X denotes the number of defective refills out of 100 inspected each day.

Table 5 : Refill inspection data

S. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total
Date	15	16	17	18	19	20	22	23	24	26	27	29	30	31	
X	9	5	6	7	6	5	6	8	7	4	6	7	6	2	84

When the process is stable, it is reasonable to assume that (i) the probability of any refill being defective is same for all refills, and (ii) the event that any of the refills being good or defective does not influence the quality of other refills.

Let

$$X_i = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ refill inspected is defective} \\ 0, & \text{otherwise} \end{cases} \quad (i = 1, 2, \dots, n),$$

where n is the sample size. For example, $n = 100$ in our situation. Then, the total number of defective refills in the sample is given by $X = X_1 + X_2 + \dots + X_n$.

Under the assumptions (i) and (ii) above, X has a binomial distribution with parameters n and p , where p is the common probability of any refill being defective. Also, we know that the mean and standard deviation (SD) of a binomial distribution with parameters n and p are given by

$$\text{Mean}(X) = np \text{ and } \text{SD}(X) = \sqrt{np(1-p)}. \quad (4)$$

In order to control the proportion of defective items produced by a process, we use a p-chart. In a p-chart, we plot the y values against the corresponding sample numbers. Unlike \bar{x} -R charts, attribute control charts have only one chart to be plotted. To arrive at the control limits, we need the mean and standard deviation of y . These are given by

$$\text{Mean}(y) = \frac{\text{Mean}(X)}{n} = p \text{ and } \text{SD}(y) = \frac{\text{SD}(X)}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}. \quad (5)$$

Recall, p is the proportion of defective items produced by the process. In quality control terminology, p is referred to as the **process average**. An estimate of p is given by

$$\bar{p} = \frac{\text{total number of defective items in all the samples}}{\text{total number of items inspected}}$$

Thus, if we have m samples, then

$$\bar{p} = \frac{\sum_{i=1}^m d_i}{\sum_{i=1}^m n_i}, \quad (6)$$

where d_i is the number of defective items in the i^{th} sample and n_i is the i^{th} sample size.

As before, the control limits in an attribute control chart are based on the 3σ limits concept. When all the subgroups are of same size ($= n$, say), then the control limits for a p-chart are given by

$$CL = p, LCL = p - 3\sqrt{\frac{p(1-p)}{n}} \text{ and } UCL = p + 3\sqrt{\frac{p(1-p)}{n}}. \quad (7)$$

Since we do not know the value of p , the control limits will be obtained by replacing p by its estimate \bar{p} given by Eqn.(6). Thus, the control limits for a p-chart are given by

$$CL = \bar{p}, LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \text{ and } UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}. \quad (8)$$

Using data from Table 5, the control limits for the refill length problem are given by

$$CL = \bar{p} = \frac{9 + 5 + \dots + 2}{14 \times 100} = \frac{84}{14 \times 100} = 0.06, \quad (9)$$

$$LCL = 0.06 - 3\sqrt{\frac{0.06(1-0.06)}{100}} = -0.011, \text{ and}$$

$$UCL = 0.06 + 3\sqrt{\frac{0.06(1-0.06)}{100}} = 0.131.$$

When the lower control limit happens to be negative, the control line is set at zero. So, $LCL = 0$ in the p-chart of refill length problem (see Fig. 7).

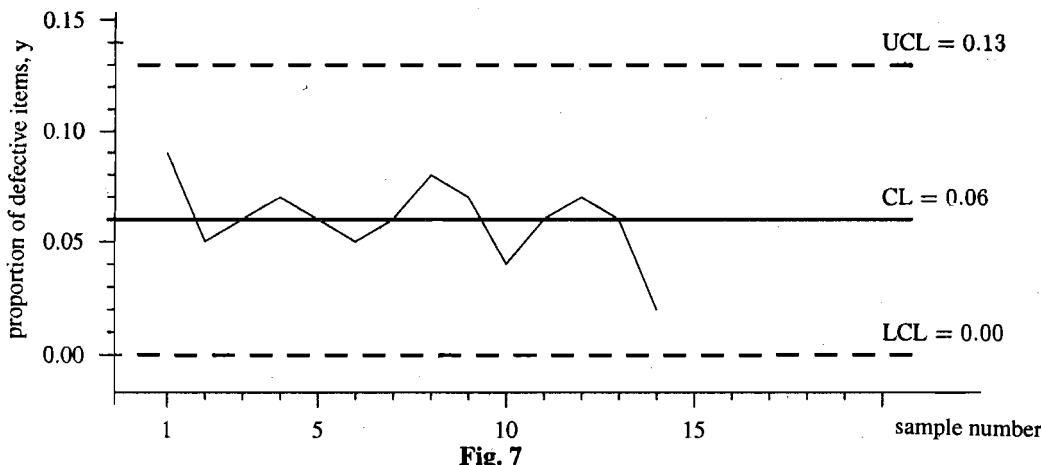


Fig. 7

We have to be cautious in interpreting a p-chart. The points which fall above the UCL are called the **high spots** and the points that fall below the LCL are called the **low spots**.

A *high spot* may be due to deterioration of the process or it could be due to a change in the inspection standard (more stringent inspection may result in larger number of defective items). Similarly, a *low spot* may indicate an improvement in the process or a deterioration in the inspection standards.

Try the following exercise.

- E6) Let the refill inspection data collected be as given in the following table.

Table 6 : Refill inspection data

S.No	1	2	3	4	5	6	7	8	9	10	Total
Date	6	7	8	9	10	12	13	14	15	16	
X	1	3	2	2	1	0	5	1	1	3	19

Recall, X denotes the *number of defective refills* out of 100 inspected each day.

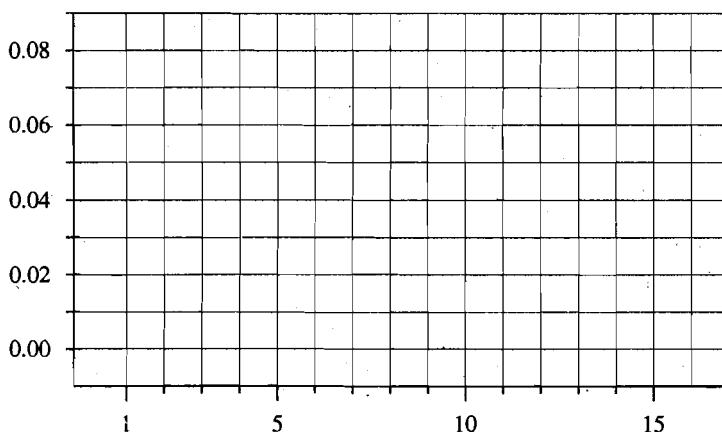


Fig. 8

- (a) Plot the first five samples in Fig. 7 itself (use a pencil).
- (b) Do you notice any change in the process? If so, what is the difference?
- (c) Can you say, from the chart you plotted, that the process is out of control?
- (d) Construct a new p-chart using only the data given in Table 6 and comment on the process (use the blank chart given in Fig. 8 above).
- (e) Was there any improvement in the process? Estimate the rejection percentages for the two periods.

Now, suppose we are not inspecting the same number of refills each day. Then, in this case, the subgroup sample sizes are varying. In such cases, the control limits will vary depending upon the subgroup sample size. Let us say, we have m subgroups with n_i sample items for the i th subgroup, then the control limits for the subgroups are given by

$$LCL_i = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} \text{ and } UCL_i = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}. \quad (10)$$

Varying Sample Sizes

The CL is drawn at \bar{p} and there is no change in the formula for \bar{p} (Eqn.(6) takes care of the unequal sample sizes).

You try the following exercise now.

E7) Construct the control limits of p-chart for the following data.

Table 7 : Data for p-chart with unequal sample sizes

S. No.	1	2	3	4	5
sample size	100	121	81	100	121
number of defective pens	2	2	0	1	2

Draw a rough sketch of the p-chart for this data.

np - charts

Sometimes it is necessary (or convenient) to look at the *number of defective items* rather than the *proportion of defective items*. In such situations, we use np - charts instead of p - charts. The only difference between p - charts and np - charts is that in the later case *y-axis represents the number of defective items in a subgroup*.

Try to convince yourself that the control limits on an np-chart should be

$$CL = n\bar{p}, LCL = n\bar{p} - 3\sqrt{n\bar{p}(1 - \bar{p})}, \text{ and } UCL = n\bar{p} + 3\sqrt{n\bar{p}(1 - \bar{p})},$$

where \bar{p} is same as defined for p-charts above.

Try the following exercise now.

E8) If the subgroup sample sizes are not the same, will the control limits vary in an np-chart? In particular, what can you say about the center line, CL?

c - chart

Finally, let us discuss the *control charts, which are used to find the number of defects per unit*. These charts are useful in situations when the performance of a process is assessed by number of defects per unit. Note that a unit may be a single product or a fixed number of products.

PCBs are used in many electronic products such as TV, Computer, etc.

For example, a printed circuit board (PCB) having several hundreds of circuits built in it, may be treated as a unit. On the other hand, a bunch of ball pens packed in a carton may also be treated as a unit.

The quality characteristic plotted on a c-chart is the total number of defects per unit, denoted by c. Usually, Poisson distribution is a good approximation for c. Therefore, a c-chart is constructed based on the assumption that c follows Poisson distribution. We know that if m is the mean of a Poisson distribution, then its standard deviation is equal to \sqrt{m} .

Hence, the *control limits for a c-chart* are given by

$$CL = m, LCL = m - 3\sqrt{m}, \text{ and } UCL = m + 3\sqrt{m}.$$

As in the case of p-charts, the control limits are obtained by replacing m by its estimate \bar{m} in the above formulae. If k is the total number of units inspected, then the average number of defects per unit, denoted by \bar{m} , is given by

$$\bar{m} = \frac{\text{total number of defects in } k \text{ units}}{k}.$$

The following problem will explain what a c-chart is and how it is applied.

Problem 3. The assembly section of M/s BP & Co. has five groups of operators. Each group consists of 3 operators. The job of the groups is to assemble various components into ball pens and pack them in cartons. The groups are also responsible for identifying and setting aside the defective components while assembling. Each carton consists of 200 pens. A sample of one carton from each group is selected at random and all the pens in the carton are inspected. The total number of defects per carton (c) is recorded

for each group. The performance of each group is monitored by maintaining a *c*-chart for each group separately. For one of the group, find (i) the *average number of defects per unit*, (ii) the *control limits for c-chart*, and (iii) *plot the control chart*.

Solution. We plot the *c*-chart for group A of BP & Co. using the data as given in Table 8.

Table 8 : Assembly defects data for group A.

S. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
date	6	7	8	9	10	12	13	14	15	16	17	18	19	20	22	
c	3	3	3	5	4	4	6	1	10	4	11	7	3	5	3	72

Here, the *average number of defects per carton* is given by

$$\bar{m} = \frac{3 + 3 + \dots + 5 + 3}{15} = \frac{72}{15} = 4.8.$$

And, the *control limits for c-chart* are given by

$$CL = 14.8, LCL = 4.8 - 3\sqrt{4.8} = -1.77 \text{ and } UCL = 4.8 + 3\sqrt{4.8} = 11.37.$$

Since we are plotting the *number of defects per carton* (*c*) on the chart, the control limits are drawn using *LCL* = 0 and *UCL* = 11.37. The corresponding control chart is as shown in Fig. 9.

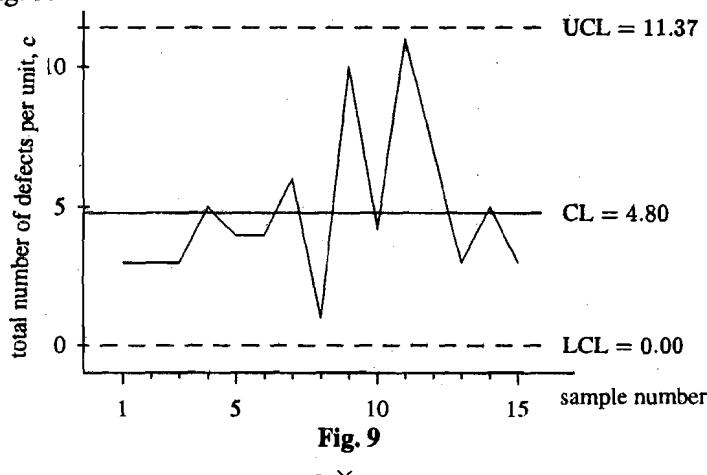


Fig. 9

Try the following exercise to have a comparison between groups A and E.

E9) The table below gives the defects for group E for the same period.

Table 9 : Assembly defects data for group E

S. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
Date	6	7	8	9	10	12	13	14	15	16	17	18	19	20	22	
c	9	10	12	13	12	8	10	12	14	2	8	10	11	10	9	150

- (a) Estimate the average number of defects per carton.
- (b) Is the assembly process of group E under statistical control?
- (c) Which of the groups A and E, do you think is better?
- (d) Explain how displaying the *c*-charts in front of these groups will help in improving the process.

In this section, we have discussed use of control charts in building quality into a product while it is being produced. Next, we discuss with you the concept of *acceptance sampling* – a techniques used to ensure that the produced products conform to specified quality standards. Here, it is important to remember that process controls are used to control and systematically improve quality, but acceptance sampling is not.

11.4 ACCEPTANCE SAMPLING

You know that, in most situations, we don't buy products directly from their

Inspection is the process of comparing actual measurable characteristics with pre-determined standard characteristics. And, 100% inspection means inspecting all items in the lot.

It is assumed that there are no *inspection errors* i.e., an item, on inspection, found to be defective if and only if it is actually so.

The maximum allowable number of defective units in a sample is called the acceptance number, which we denote by c .

manufacturer. The products from the manufacturer are first supplied to dealers, the dealers supply them to retail shops and we buy them from retail shops.

Let us consider the case of Mr. Anil who is one of the dealers for M/s BP & Company. Mr. Anil buys ball pens from the company in large quantities. He receives the pens in lots where each lot contains 1000 pens. Mr. Anil will be too happy if all the lots that he receives have no defective pens at all. But, we know that this never happens in practice.

Can we then think of a procedure that will ensure that no lot has a defective pen? If so, at what cost? Is it worth adopting such a procedure? You might think that 100% inspection before packing is a procedure that ensures *defect-free* lots. In fact, 100% inspection is not always 100% efficient.

In some cases, there may be slips in inspection due to fatigue or measuring equipment errors. And, in some other situations, it may not be feasible to carry out 100% inspection. For example, if the product is bullets and inspection involves firing the bullets, then 100% inspection means, we will be left with nothing. So, what is the alternative?

We use sampling techniques in most such situations. In fact, sampling is extensively used in our day to day life. For example, we use sampling while purchasing vegetables and groceries, for selection of our family doctor, and so on.

A **sampling procedure** used to accept or reject a lot of items is known as an *acceptance sampling plan* (**ASP**, in short). Indeed, we use *acceptance sampling plan* as an audit tool to ensure that the product of a process conforms to requirements.

11.4.1 Sampling Plan Concepts

For a smooth discussion of the topic, we need an understanding of terms defined in the following definition.

Definition. A **lot** is a collection of units of product picked for the purpose of sampling. Based on the results of inspection of a random sample from the lot a decision is made to accept or reject all the units in the lot. The act of accepting or rejecting the entire lot is called **sentencing the lot**.

The number of units in a lot is called the **lot size**. The number of units inspected to sentence a lot is called the **sample size**. The proportion of defective items in a lot is called the **lot quality**. Throughout, we shall use (i) N for the *lot size*, (ii) n for its *sample size*, and (iii) p for the *lot quality*.

There are three approaches to *lot sentencing*: (1) no inspection; (2) 100% inspection; and (3) acceptance sampling. Here, we have to discuss the acceptance sampling and, to understand this in a better way, let us start with a simple example of an acceptance sampling plan for Mr. Anil, which we denote by **ASP1**.

ASP1. From each lot of 1000 pens, take 100 at random and inspect them. Accept the lot if the inspected sample contains at most one defective pen; otherwise reject it.

Thus, for **ASP1**, $N = 1000$ and $n = 100$. So, if a lot consists of 20 defective pens, then the lot quality $p = 0.02$. Also, observe that acceptance number $c = 1$ in this case.

Now identify these quantities in the following exercise.

E10) Specify N , n and p for

- (a) a lot of 400 bolts having 36 defective bolts;
 - (b) a box of 50 cricket balls having two defective balls;
 - (c) a situation when a sample of 40 bolts is drawn from a lot of 400 bolts and the lot is rejected if the sample contains two or more defectives.
-

For sentencing, suppose a lot is subjected to ASP1 by Mr. Anil. What do you think is the probability that the lot will be accepted? Of course, this is *one*, if all the pens in the lot are good; and it is *zero*, if all the pens in the lot are defective. Thus, the probability of accepting a lot, denoted by P_a , depends on the number of defective items in the lot.

Suppose a lot has 20 defective items under ASP1. What is P_a for such a lot? Since *ASP1 allows at most one defective pen* in a sample of 100 pens, this probability is given by

$$P_a = P[X = 0] + P[X = 1],$$

where X is the number of defective pens in the sample.

We know that, in general, X has hypergeometric distribution and, so, we can compute the above probabilities using this fact. However, we know that when lot size N is large compared to sample size n , these probabilities can be closely approximated by assuming that X follows binomial distribution with parameters n and p .

In fact, when $N \geq 10n$, it hardly makes any difference whether the probabilities are computed using *hypergeometric distribution* or *binomial approximation*. The advantage of using binomial approximation is that the formulae are simple and the numbers are small.

In general, the (binomial) *probability of observing exactly k defects in a sample of size n* is given by

$$P[X = k] = {}^n C_k p^k (1 - p)^{n-k},$$

where ${}^n C_k$ stands for the number of ways of choosing k items out of n items. And, the probability of acceptance (P_a) is given by

$$P_a = P[X \leq k] = \sum_{d=0}^k {}^n C_d p^d (1 - p)^{n-d} \quad (11)$$

Thus, for $n = 100$ and $p = 0.02$, we get $P_a = P[X \leq 1] = 0.1326 + 0.2706 = 0.4032$.

Note that this means that if a number of lots with $p = 0.02$ are subjected to ASP1, then only about 40.32% of them will be accepted and about 59.68% of the lots get rejected, even though their quality is same as those accepted.

Since P_a depends on p , from now onwards, we shall write it more explicitly as $P_a(p)$. Then, by above calculations, $P_a(0.02) = 0.4032$, for ASP1.

Try the following exercise.

E11) For ASP1, compute (a) $P_a(0.03)$; and (b) $P_a(0.05)$.

The three curves shown in Fig. 10 (for $c = 0, 1$, and 2) are developed by evaluating Eqn.(11) for various values of p , $0 \leq p \leq 1$. So, each point on a curve is represented by $(p, P_a(p))$.

Each curve is an **oc curve** (*operating characteristic curve*) for some acceptance sampling plan. The preference of an acceptance sampling plan is completely described by its *oc curve*.

In view of Eqn.(11), a typical *oc curve* is a pictorial representation of the relationship between the *lot quality* (p) and the *probability of acceptance* (P_a), for a given sampling plan. The greater the slope of an *oc curve*, the greater is the discriminatory power. However, as c decreases, the *oc curve* gets shifted to left (without much a change in its slope).

In general, the exact shape of a specific *oc curve* depends on the values of parameters

Hypergeometric distribution with parameters N , G and n is the distribution of the number of *good* objects in a simple random sample of size n from a population of N objects of which G are *good*.

AQL, LTPD, α and β . Very shortly, we will define these four terms in this section.

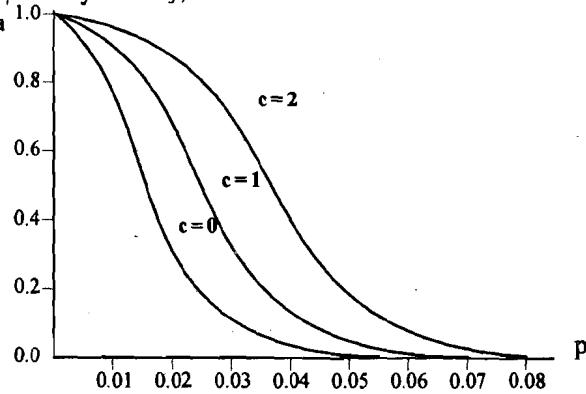


Fig. 10. *oc curves* for different values of c .

Here, we ask: What will happen if we change sampling plan ASP1? Let us consider another acceptance sampling plan with $c = 3$, which we denote by ASP2.

ASP2. From each lot of 1000 pens, inspect 100 at random and accept the lot if the inspected sample has at most three defective pens.

Once again, let us compute P_a for a lot which has exactly 20 defective items (i.e., $p = 0.02$). Under ASP2, it is given by

$$P_a(0.02) = \sum_{d=0}^3 {}^n C_d p^d (1-p)^{n-d} = 0.86.$$

On comparing, we find that a lot of the *same quality* (as $p = 0.02$, in each case) has a chance of 0.40 (rounded off to two decimal points) of getting accepted, if it is subjected to ASP1, and has a chance of 0.86, if it is subjected to ASP2. So, we conclude that the probability of accepting a lot depends on both (a) the lot quality; and (b) the *acceptance sampling plan*.

Just as for ASP1, we can have *oc curve* for ASP2. In Fig. 11, the *oc curves* for three different acceptance plans are shown for comparison.

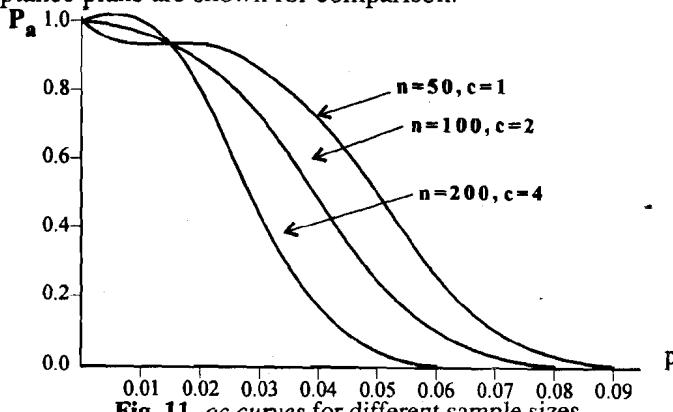


Fig. 11. *oc curves* for different sample sizes.

In Table 10, we have listed P_a values under ASP1 and ASP2 for some selected values of p . Examine these values to compare the two sampling plans.

Table 10 : Lot acceptance probabilities, P_a

P	P_a under ASP1	P_a under ASP2
0.000	1.0000	1.0000
0.010	0.7358	0.9816
0.020	0.4033	0.8590
0.030	0.1946	0.6472
0.046	0.0525	0.3196
0.076	0.0034	0.0490
0.100	0.0003	0.0078

We shall continue discussing *oc curve* a little later. Here, we take a break to talk about certain important parameters, which are used while adopting an acceptance sampling plan.

Both, M/s BP & Co. and Mr. Anil, understand that supply of completely defect-free lots is not possible. So, they have to come to a compromise. This is how a bargain starts between the two. Finally, they reach to an agreement : *Mr. Anil will accept majority of lots which have at most 1% defective pens (i.e., $p = 0.01$) and the company will take back all those lots which contain more than 1% defective pens.*

A level of quality, which is mutually agreed upon by both the buyer and the seller is called **acceptable quality level (AQL)**. Thus, in view of above agreement between Mr. Anil and M/s BP & Co., $AQL = 0.01 (= p)$. This means that Mr. Anil should accept lots with $p \leq 0.01$ in *majority* of the cases.

Now, let us examine the consequence of Mr. Anil's decision, when $AQL = 0.01$ and ASP1 is adopted by him.

From Table 10, we find that $P_a(0.01) = 0.7358$. So, under ASP1, Mr. Anil will reject 26.42% ($= 100(1 - 0.7358)$) of the lots whose quality is 0.01. This is obviously a *risk* to the producer because it was agreed upon by both that lots of quality 0.01 will be accepted in majority of the cases whereas Mr. Anil is rejecting 26.42% of them. Of course, a lot with $p < 0.01$ has a smaller chance of getting rejected than 26.42% (see Table 10). In other words, the producer's risk for any $p < AQL$ is less than 26.42%.

So, here is an important observation to note: *Among all values of p between 0 and AQL , producer's risk is maximum when $p = AQL$.* Producer's risk is defined as the probability of rejecting a lot whose $p = AQL$ and is denoted by α . Usually it is expressed as a percentage. So, producer's risk in above situation is given by $\alpha = 100(1 - P_a(AQL))\%$.

What will happen if Mr. Anil adopts ASP2 instead of ASP1? Try the following exercise to find the answer.

E12) Assuming ASP2 is adopted, find out $P_a(0.01)$. Also locate (roughly) the points $AQL (= 0.01)$ on x-axis, producer's risk α on y-axis and $(AQL, P_a(AQL))$ on the two *oc curves* plotted under ASP1 and ASP2.

With $AQL = 0.01$ in above exercise, you must have got the producer's risk as 1.84% under ASP2. Thus, under ASP1, the producer's risk is more and, so, the company would say to Anil: *26.42% is too much of a risk for us, we would bear at most 5% risk.* What is the solution? We can change the sampling plan.

On the other hand, under ASP2, the producer's risk is only 1.84% (much less than the desired 5%). So, the company will be complacent. But then, what will happen to Mr. Anil as a consumer?

From Eqn.(11), we find that $P_a(0.05) = 0.2578$, under ASP2. This means that, if ASP2 is adopted, then lots whose quality is as bad as having 5% defective pens (this is much worse than the agreed upon quality $AQL = 0.01$) will get accepted 25.78% of the time. Obviously, this is a risk to the consumer.

Why should Mr. Anil accept such bad lots 25.78% of the time? Accepting lots whose quality is worse than AQL, the consumer is at a loss. Thus, ASP2 is not good for Mr. Anil even though it is good for the producer.

If we want a sampling plan that is best for both the consumer and the producer, then all those lots with $p \leq AQL$ should be accepted with probability 1 and all those lots with $p > QAL$ should be rejected with probability 1 (or accepted with probability 0).

The *oc curve* of such a sampling plan will look like the one in Fig. 12 and is called an **ideal oc curve**. It is clear that such an acceptance sampling plan would call for almost

Acceptable Quality Level

Producer's Risk

Consumer's Risk

100% inspection and the inspection costs will be high.

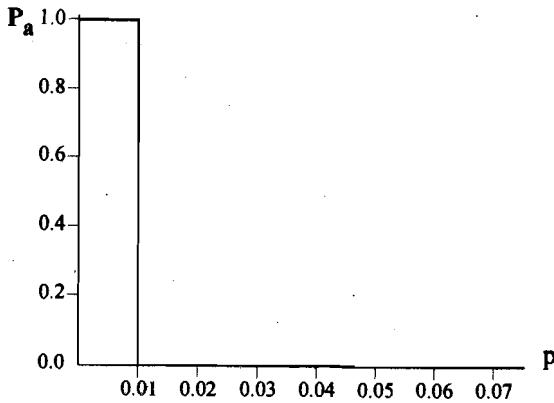


Fig. 12. Ideal oc curve.

Lot Tolerance Proportion Defective

As a compromise, Mr. Anil is ready to tolerate lot qualities worse than AQL upto certain limit, but not beyond. Let us assume Mr. Anil is willing to tolerate lots with $p \geq 0.05$ but not more than 10% of the time. The *tolerance limit* on the lot quality $p = 0.05$ that he has chosen to tolerate, is called the **lot tolerance proportion defective** and is denoted by **LTPD**. Thus, $LTPD = 0.05$ in case of Mr. Anil.

In the above paragraph, we have mentioned that Mr. Anil is willing to accept lots with quality $p = LTPD$ only 10% of the time. Here, 10% is what we call the consumer's risk.

The probability of accepting a lot with $p = LTPD$ is called the consumer's risk and is denoted by β . In other words, *consumer's risk* is equal to $P_a(LTPD)$. Like the producer's risk, the consumer's risk is also usually expressed as a percentage.

Above we have seen that while ASP1 is good for the consumer, ASP2 is good for the producer. But neither of the two plans will satisfy both of them. So, we should look out for a sampling plan that should be acceptable to both consumer and producer.

It is customary to use the AQL and LTPD points for this purpose and the corresponding points on the *oc curve*, α and β , respectively, give producer's and consumer's risk (see Fig. 13). Such an acceptance sampling plan protects the interest of both producer and consumer.

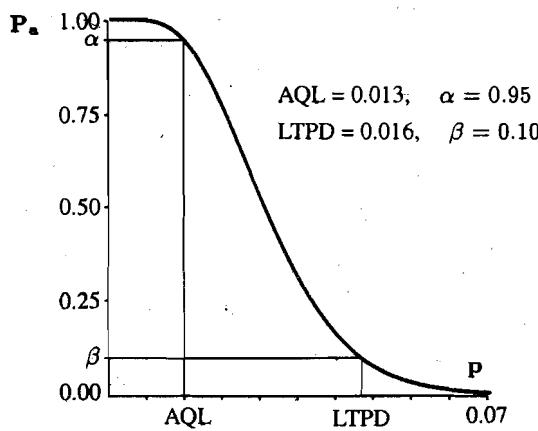


Fig. 13

Note that P_a for any lot with $p < AQL$ is at least $1 - \alpha$ (or probability of rejecting such a lot is at most α) and P_a for any lot with $p > LTPD$ is at most β .

Obtaining acceptance sampling plans that protect both producer and consumer usually calls for large sample sizes. So, in practice, one has to compromise somewhere.

A sampling plan never guarantees the acceptance of 100% perfect material. In fact, some defective material will also get accepted in the process. But, by proper specification of β risk, it is possible to minimise the amount of defective material.

If you want to have cake and eat it too, then you must have too many cakes.

Thus, a sampling plans is used to have a reasonably good idea about how much unacceptable material will be involved.

11.4.2 Single Sampling Plans

We use some statistical and probability tools to develop sampling plans that meet the desired α & β risks and maintain the desired AQL and LTPD quality levels. Of course, our main objective is to determine P_a of lots with varying quality.

There are many types of acceptance sampling plans. As before, we will confine our discussion to *single sampling plans* and learn how to design them. Other types of sampling plans will be discussed very briefly at the end of the section.

Already, we discussed some examples like *ASP1* and *ASP2*. More generally, a *single sampling plan* is completely described by specifying (i) the lot size, N ; (ii) sample size, n ; and (iii) acceptance number, c .

Single Sampling Plans

In a *single sampling plan*, given by the three values (N, n, c), we inspect n items at random from a lot of N items and accept the lot if the number of defective items in the sample is less than or equal to the acceptance number c ; otherwise we reject the lot. And, throughout, we use *simple random sampling without replacement* for sampling items.

A simple graphical procedure, called a *binomial nomograph*, is used to construct single sampling plans for a specified AQL, LTPD, α and β (see Fig. 14).

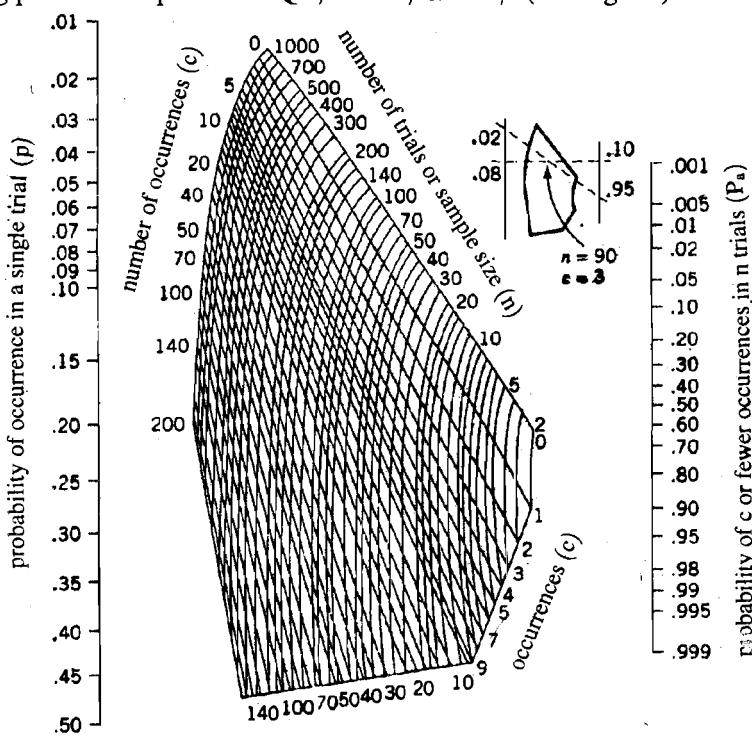


Fig. 14. (Source: D.C.Montgomery, Introduction to Statistical Quality Control (4/e),2001.)

As an illustration, we try to find a single sampling plan when AQL = 0.02, LTPD = 0.08, $\alpha = 0.05$ and $\beta = 0.10$.

Firstly, draw a straight line joining 0.02 on the p-scale and 0.95 ($= 1 - \alpha$) on the P_a -scale. Draw another straight line joining 0.08 on the p-scale and 0.1 ($= \beta$) on the P_a -scale. Now, read the values of n and c corresponding to the point of intersection of the two lines drawn. You can see that we will get $n = 90$ and $c = 3$. Observe that we have not talked about lot size anywhere in the process.

Infact, the lot size is implicitly assumed to be at least 10 times the sample size. Thus, if

Mr. Anil wants to use the above derived plan, his lot size should be at least 900. And, as $N = 1000$ for refill length problem, the above plan can be used.

Try the following exercise.

E13) From the nomograph shown in Fig. 14, derive single sampling plans when

- $AQL = 0.01$, $LTPD = 0.10$, $\alpha = 0.05$, $\beta = 0.10$; and
- $AQL = 0.03$, $LTPD = 0.08$, $\alpha = 0.05$, $\beta = 0.10$.

Percentage Sampling

In the past, it was a common practice in industry to inspect certain percentage of items in the lots. In other words, single sampling plans of the type $(100, 10, 0)$, $(500, 50, 0)$ and $(1000, 100, 0)$ were often being used. The sample size in all these plans is 10% of the lot size.

This may make us believe that all these plans are equally good. No, they are not! On examining the *oc curves* for the 3 plans mentioned in this paragraph with $c = 0$ (see Fig. 15), it is sufficiently clearly that the three plans are drastically different.

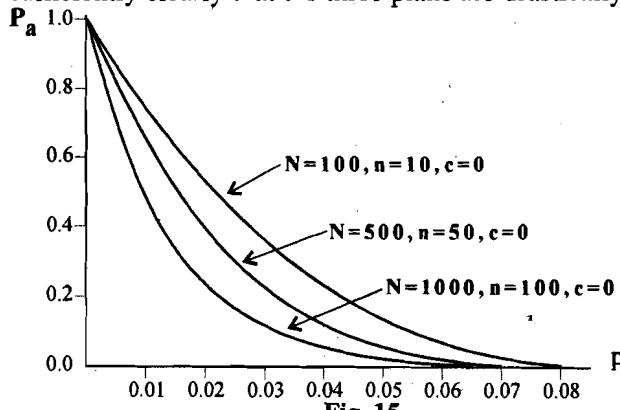


Fig. 15

The main disadvantage in this approach is that *different sample sizes offer different levels of protection*. As such, it is illogical for the level of protection the consumer enjoys for a critical part to vary as the size of the lot varies.

You will find the following exercise more convincing.

E14) Assuming $AQL = 0.05$, read, roughly, the producer's risks for the 3 plans in Fig. 15.

There are many other sampling plans such as *double sampling plans*, *multiple sampling plans*, *sequential sampling plans*, *chain sampling plans*, and so on. A number of published sampling plans are developed from various view points. Of course, many computer softwares are available today in the market with which you can design and evaluate various sampling plans and schemes at your finger tips.

With this we have come to the end of the block. Let us summarise what we have learnt in this unit.

11.5 SUMMARY

In this unit, we have discussed with you the following aspects of SQC.

1. The concept of quality and its characteristics.
2. How quality of a product is achieved by ensuring that quality characteristics conform to their specifications.
3. Use of control charts as a primary tool for an on-line process control.
4. The construction and application of \bar{x} -R charts for controlling variable characteristics.

5. The meaning of process capability and its evaluation through \bar{x} -R charts.
6. The construction and application of p, np, and c charts for controlling attribute quality characteristics.
7. The concept of acceptance sampling plans and use of the OC curve, AQL, producer's risk, consumer's risk, and LTPD, while arriving at a plan that protect the interest of both consumer and producer's equally. Also, we have seen how OC curves have been used in comparing sampling plans.
8. The construction of a single sampling plan using *binomial nomograph*.

11.6 SOLUTION/ANSWERS

E1) Measurable characteristics such as diameter of a ball, length of refill, weight of cricket ball, thickness of a washer etc., are suitable for control charts for variables. Characteristics such as defective ball pens, defects such as scratches on a cricket ball, neps and faded portions on a piece of cloth, etc., are suitable for control charts for attributes.

E2) Estimate of process mean is given by

$$\bar{\bar{x}} = \frac{\text{sum of sample averages}}{\text{number of samples}} = \frac{7575}{25} = 303 \text{ grams.}$$

Estimate of process standard deviation is given by

$$\bar{R} = \frac{\text{sum of sample ranges/number of samples}}{d_2} = \frac{154/25}{2.059} = 2.992 \text{ grams.}$$

Then, the *control limits for \bar{x} - chart* are given by

$$CL = 303, LCL = \bar{\bar{x}} - A_2\bar{R} = 303 - 0.73 \times 6.16 = 298.503,$$

$$UCL = 303 + 0.73 \times 6.16 = 307.497.$$

And, the *control limits for R-chart* are given by

$$CL = \bar{R} = 6.16, LCL = d_3\bar{R} = 0, UCL = d_4\bar{R} = 2.28 \times 6.16 = 14.045.$$

E3) After dropping 12th subgroup,

$$\bar{\bar{x}} = \frac{201.808 - 10.086}{19} = 10.091, \bar{R} = \frac{3.51 - 0.38}{19} = 0.1647.$$

Therefore the revised control limits, for \bar{x} - chart, are

$$LCL = 9.995, CL = 10.091, UCL = 10.186; \text{ and}$$

$$LCL = 0, CL = 0.1647, UCL = 0.3476, \text{ for R - chart.}$$

From Table 1, we find that none of the points (excluding 12th subgroup), either on \bar{x} - chart or R - chart, fall outside control limits. So, the remaining data indicate statistical control.

E4) We may take refill length problem. Here, total tolerance ($= T$) = 0.4. Then,

$$C_p = \frac{0.4}{6\sigma} = 1.5 \Rightarrow \sigma = \frac{1.6}{36} = 0.044.$$

Thus, (a) rejections will occur on USL side when $\mu + 3\sigma > USL$. For instance, if $\mu = 10.111$ (i.e., $\mu + 2\sigma = USL$), there will be rejections on USL side. And, (b) rejections will occur on LSL side when $\mu + 3\sigma < LSL$. For instance, if $\mu = 9.889$ (i.e., $\mu - 2\sigma = LSL$), there will be rejections on LSL side.

E5) For the combined data, $\bar{\bar{x}} = \frac{200.177}{20} = 10.009$ and $\bar{R} = \frac{2.65}{20} = 0.1325$. Thus, (a) estimates of mean and standard deviation for the new process are given by

$$\hat{\mu} = \bar{\bar{x}} = 10.009 \text{ and } \hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{0.1325}{2.326} = 0.0569.$$

And, (b) the control limits, for \bar{x} - chart, are given by

$$LCL = 9.932, CL = 10.009, UCL = 10.086; \text{ and}$$

$$LCL = 0, CL = 0.1325, UCL = 0.2795, \text{ for R - chart.}$$

Again, (c) the first subgroup average is below LCL. on \bar{x} - chart. Possibly there was an assignable cause. But for this, the data indicate statistical control. And, (d) since R-chart indicates control, we might use all the 20 subgroups to estimate the process variability. An estimate of the process capability is given by

$$6\hat{\sigma} = 6 \times \frac{\bar{R}}{d_2} = 6 \times \frac{0.1325}{2.326} = 0.3414,$$

and, an estimate of process capability ratio is given by

$$\hat{C}_p = \frac{T}{6\hat{\sigma}} = \frac{0.4}{0.3414} = 1.171.$$

Finally, (e) since $\hat{C}_p < 1.33$, further reduction in process variability is essential. So, the advice is that the manager should explore the possibilities of reducing variation further.

- E6) (b) There is clear indication that the process average has shifted downwards. (c) We cannot comment on the process control unless we redraw the control limits.

(d) For the data in Table 6, $CL = \bar{p} = \frac{19}{10 \times 100} = 0.019$,

$$LCL = 0.019 - 3\sqrt{\frac{0.019 \times 0.981}{100}} = -0.02, UCL = 0.019 + 3\sqrt{\frac{0.019 \times 0.981}{100}} = 0.06.$$

Since, $LCL < 0$, LCL is plotted at 0.0. (e) Data indicate improvement in the process. The rejection percentage for the first period is equal to 6% ($= 100 \times$ old \bar{p} estimate) and for the latter period it is equal to 1.9% ($= 100 \times$ new \bar{p} estimate).

- E7) The estimate of process average is given by $\bar{p} = \frac{7}{253} = 0.013$. In this case, the LCL turns out to be zero for all the five subgroups. The UCL for subgroups with subgroup size 100 is equal to $0.013 + 3\sqrt{\frac{0.013 \times 0.987}{100}} = 0.048$. The table below summarises the UCLs and sample averages for the data.

S.No.	1	2	3	4	5
sample size	100	121	81	100	121
sample average	0.020	0.016	0.000	0.010	0.016
UCL	0.048	0.045	0.052	0.048	0.045

- E8) YES. The control limits will vary with varying sample sizes. Even the center line will vary because $CL = np$.

- E9) (a) Estimated number of defects per carton $= \frac{150}{15} = 10$. (b) The control limits are : $CL = 10$, $LCL = 10 - 3\sqrt{10} = 0.513$ and $UCL = 19.487$. Since what we plot on the y-axis is the total number of defects per carton, we may take $LCL = 1$ and $UCL = 19$. Clearly, assembly process of Group E is under statistical control. (c) Since the average number of defects per carton of Group A is only 4.8, Group A is better. (d) If the chart is displayed in front of the operators, they will have a continuous feed back on their performance. So, whenever the quality deteriorates, they can correct themselves. Psychologically, it will have good impact on the operators.

- E10) (a) $N = 400, p = 0.09$; (b) $N = 50, p = 0.04$; (c) $N = 400, n = 40, p = 0.005$.

- E11) (a) $P_a(0.03) = 0.1946$; (b) $P_a(0.05) = 0.0371$.

- E12) Under ASP2, $P_a(0.01) = 0.9816$ (see Table 10). The actual points are $(0.01, 0.7358)$, for ASP1, and $(0.01, 0.9816)$, for ASP2. So, these should be the points that you read from the graph approximately.

- E13) The single sampling plans from the nomograph are (i) $(N, 40, 1)$ and (ii) $(N, 200, 10)$. Here, the lot size N should be at least 10 times the corresponding sample size.

- E14) The producer's risks are: 0.2141, for $(1000, 100, 0)$; 0.2578, for $(500, 50, 0)$; and 0.1498, for $(100, 10, 0)$.

APPENDIX

TABLE-1
F-distribution

$F_{0.05}$

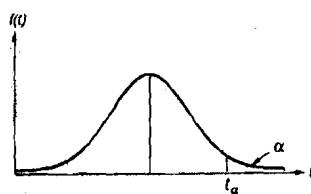
v_2 = Degrees of freedom for denominator	v_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.50	19.00	19.20	19.20	19.30	19.30	19.40	19.40	19.40	19.40	19.40	19.40	19.40	19.50	19.50	19.50	19.50	19.50	19.50
3	10.10	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.38	2.38	2.30	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	3.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.93
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

TABLE-1(Continued)
F-distribution

$F_{0.01}$

ν_2 = Degrees of freedom for denominator		ν_1 = Degrees of freedom for numerator																
1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4.052	5.000	5.403	5.625	5.764	5.859	5.928	5.982	6.023	6.056	6.106	6.157	6.209	6.235	6.261	6.287	6.313	6.366
2	98.50	99.00	99.20	99.30	99.40	99.40	99.40	99.40	99.40	99.40	99.40	99.40	99.50	99.50	99.50	99.50	99.50	99.50
3	34.10	30.80	29.50	28.70	27.90	27.70	27.50	27.30	27.20	27.10	26.90	26.70	26.60	26.50	26.40	26.30	26.30	26.10
4	21.20	18.00	16.70	16.00	15.50	15.20	15.00	14.80	14.70	14.50	14.40	14.20	14.10	13.90	13.80	13.70	13.60	13.50
5	16.30	13.30	12.10	11.40	10.70	10.30	10.20	10.10	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	
6	13.70	10.90	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97
7	12.20	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74
8	11.30	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95
9	10.60	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40
10	10.00	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.06	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.01
40	7.31	5.18	4.31	3.81	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.90
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.90	1.79	1.70	1.59	1.47	1.32

TABLE-2
t-distribution



<i>v</i>	<i>t_{.100}</i>	<i>t_{.050}</i>	<i>t_{.025}</i>	<i>t_{.010}</i>	<i>t_{.005}</i>	<i>t_{.001}</i>	<i>t_{.0005}</i>
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

UNIT 12 SIMPLE RANDOM SAMPLING AND SYSTEMATIC SAMPLING

Structure	Page No.
12.1 Introduction Objectives	5
12.2 Sampling – What and Why?	6
12.3 Preliminaries	7
12.4 Simple Random Sampling	9
12.5 Estimation of Population Parameters	13
12.6 Systematic Sampling Linear Systematic Sampling Circular Systematic Sampling Advantages and Limitations of Systematic Sampling.	18
12.7 Summary	21
12.8 Solutions/Answers	21
Appendix-A	

12.1 INTRODUCTION

In our day to day routine we often have to make certain judgements about a large bulk or population after studying a small portion, or a sample of it. For example, a house wife tastes a spoonful of soup to see whether a little more salt is required before it is served to guests; a quality control inspector inspects a sample from a tin of oil before passing the whole tin as acceptable quality; or a doctor takes a few drops of blood from a patient to decide if the patient has malarial infection. These are typical examples where it is not practical to examine the entire lot or population and decision making is done on the basis of sample information. Essentially this is sampling. It saves time and if you have based your judgement on judicious observations, it serves your purpose. Think of situations, when decisions are to be taken at macro levels, say, national level. Sampling has become an effective tool for generating information for policy formulation at different administrative levels. The above examples have the special phenomenon that a spoon of soup, tinfoil of oil and the blood in the patient are known to be perfectly **homogeneous** material so that every part of the material represents the material exactly. Often, however, we are not in this simple situation. For example, suppose we are interested in knowing the average weight of an adult Indian male. Obviously, it will not be satisfactory to measure just a few adult males, as not all adult males are of the same weight. They show considerable **heterogeneity**. So how do we take a part of a large heterogeneous mass to draw valid conclusions from it? However, careful considerations are needed for selection of samples and making valid inferences from these samples. In the subject of 'sampling', these considerations and criteria are developed on a scientific basis.

In this unit, we shall start by introducing the basic concepts of sampling. We shall then discuss the simplest procedure for sample collection i.e., simple random sampling (SRS). In the process you will get introduced to the concept of random numbers, simple random sampling as a method of selection, estimation as well as considerations for determination of sample size.

Further, the concept of systematic sampling, which is also a method of selection based on random sampling procedure, will be introduced alongwith the estimation approach.

Objectives

After reading this unit, you should be able to

- select samples using simple random sampling and systematic sampling procedures;
- estimate population mean, population variance, proportion, alongwith their efficiency in SRS;
- select samples using systematic sampling by linear and circular systematic procedures.

12.2 SAMPLING – WHAT AND WHY?

You have already learnt in Block 1, the concept of a statistical population. Often we are interested in studying a specified characteristics of the individuals in a finite population. For example, we may be interested in studying the annual income and size of the households in Delhi for the year 2000-2001. In this case our population is the collection of all households in Delhi during the year 2000-01 and the individuals are the households. It may be enough for our purpose to find out the average size and annual income of the house-holds in Delhi. How do we go about obtaining this information? One way is to visit each household, note down the number of members and find out from the members how much was their total income during the year 2000-01 and calculate a simple average of all the figures obtained. You can very well imagine the difficulties and expenses involved in undertaking this huge task. The task of enumerating all the households in Delhi, visiting them and noting their sizes and annual income will be termed a census of the population under study. As you may all be aware, such a census is held once in ten years and information is obtained on a large number of characteristics including the ones mentioned above. Given this complexity and the huge expenses involved, this method cannot be adopted every time we need information on a population. Here are a few more instances where we may need to collect information on a characteristic from all the individuals of a population:

- i) A telephone company may be interested in figuring out the average number of calls and average duration of a call made by the households in a locality or a city
- ii) A large fruit store may be interested in the quality of the truck load of peaches packed in crates received at the stores from a farm
- iii) A company producing electric bulbs may be interested in the average life of bulbs produced by them during a shift.

E1) In the cases mentioned above identify the population and the individuals.

You may notice that observing or inspecting all the individuals in a population can be very expensive in terms of time and money. In certain instances, like observing the life of bulbs, inspection may be destructive, so that observing all the bulbs of a population for calculating their average life is a meaningless exercise. so an alternative approach is called for. Can you think of an alternative? One alternative is to observe or measure only some individuals of a population, but estimate the average for the whole population based on the few measurements made. The idea is roughly as follows: Suppose you want to find the total number of mangoes on a tree. First, it is easy to count the number of mangoes on a "typical" branch. Multiply the number obtained by the number of branches and you get an estimate. Do you think this method will yield a good enough estimate? The method adopted in this approach is called the **method of sampling**. You may realise that in this methodology your final estimate of the total may depend on the number of fruits on the selected branch. Of

course, if all the branches happen to bear the same number of fruits, it does not matter which branch you happen to select. You may therefore want to know if all the branches bear nearly the same number of fruits or, if they differ then by how many they may differ. In order to do this, again using the principle of sampling, you may select another branch to get an idea of the difference. If you select the branches with a priori known probabilities, (for example, by ensuring that any set of two branches have the same probability of being selected in the sample as any other set), you may be able to use the theory of probability and statistical inference you have learnt in the earlier blocks to calculate the likely error in your estimate. The theory and method of sampling deal with the issue of how to select the sample, how to estimate the population total or average and how to estimate the error in observing only a sample instead of the entire population. The error that results from estimation based only on a sample of observations is called **sampling error**. In contrast, there may also be nonsampling errors while observing or measuring an individual. A census approach, while free from the sampling error may suffer from nonsampling errors. In fact, in measuring a large number of individuals, due to 'inspection fatigue' the nonsampling error may be large. The theory of sampling may allow one to estimate the sampling error but will not be able to assess the extent of nonsampling errors.

In order to have a better grasp of what we have discussed above, you may try this exercise.

-
- E2) List the advantages and disadvantages of using a sampling approach instead of a census approach for studying a characteristic.
-

You may also recall at this stage that in Block-2, you learnt about sampling distributions or the derived distribution of functions of a sample of observations. The issues studied in that block are to be distinguished from the issues we will be considering in this block. In Block-2, you had a conceptual infinite population, such as the drying time required for a formulation of paint, and you had ten realizations of the random variable 'drying time' on ten pieces of wood on which this formulation of paint was applied. These ten observations were assumed to be ten independent realizations of the same random variable having a theoretical distribution and you derived the distribution of their average and other functions. The issues to be studied in this block relate to sampling from a finite population and no theoretical distribution of the characteristic is assumed.

Before proceeding further to discuss the methods of simple random sampling, we shall introduce some concepts and definitions, which we shall be using frequently in our discussion.

12.3 PRELIMINARIES

Let us assume that we wish to find out the proportion of votes a particular party A is expected to get in an election in a particular constituency.

An **element** is a unit for which information is sought. In this example the element is a registered voter in the constituency. The study variable will be measured as one if the voter prefers to vote for the party A, otherwise, the measurement will be taken as zero.

The population as you already know, is an aggregate of elements about which the inference is to be made; the collection of all the registered voters of the constituency

Sampling

in this case constitutes the population. A population is **finite** if it consists of a finite number of elements. In this unit we shall consider the case of finite population only.

For studying a population we select some of the elements or a collection of elements (i.e. a sample) of the population on which observations are made. These elements are called the **sampling units**. In our example, if households, which has got a number of elements i.e., individual voters, are to be selected then households are sampling units. Sampling units are non-overlapping collections of elements of the population.

For selection purposes, identity of sampling units is necessary. Usually, a list of sampling units of the population provides such an identity. A complete list of sampling units which represents the population to be covered is called a **sampling frame**. The number of units in the sample is the **sample size**.

In the entire theory of sampling, the approach for selecting a **representative** sample and making a good **estimate** from the sample is addressed. In the example considered here, preference for the party is asked only from the registered voters selected in the sample. This information is then used to determine the proportion of all votes that party A is expected to get in the election. It is therefore necessary to exercise a great deal of care in selecting sampling units for a sample survey. Other examples of results based on sample surveys are quite common in practice. Whenever you read about important figures like production of important crops, or average income of people in rural or urban areas, it must be realised that such figures are invariably based on results of well planned sample surveys.

If the units in the sample are selected using some random mechanism then such a procedure is called **random sampling** or **probability sampling**. In this method samples are selected according to certain laws of probability in which each unit of the population has some definite probability of being selected in the sample. All other sampling procedures, which are not based on random procedures but are based on subjective judgement or convenience of the sampler, are known as **non-random sampling** or **non-probability sampling**. They are also termed as **purposive sampling** or **judgement sampling**. Clearly, inferences drawn on the basis of a purposive sample can often be subjective and biased.

Random sampling is preferred over non-random sampling for a variety of reasons. Besides eliminating the subjectivity in selection, it provides a measure of reliability associated with the estimates developed from the samples. Thus, one can make inferences from the sample with a known level of confidence. As stated earlier, in random sampling procedure, every unit in the population is assigned definite probability of selection. The randomness associated with the sampling procedure is the key to make valid inferences from the sample.

Samples are often selected by adopting the procedure of **one after the other draw procedure** or **unit by unit** selection. If the units selected at one draw are replaced in the population before the next draw then the procedure is called **with replacement** (WR) procedure. If the units are not replaced in the population and the selection is made from the remaining units then the selection is called **without replacement** (WOR). If a population consists of N units and a sample of size n is to be selected, number of possible samples for with replacement procedure is N^n . In case of without replacement sample, if the order of sample units is ignored then there are $\binom{N}{n}$ possible samples.

You may now try to solve the following exercises to see whether you have grasped the basic concepts of sampling discussed above.

- E3) Define population, sampling unit and sampling frame for conducting surveys on each of the following subjects.
- Measurement of the volume of timber available in a forest.
 - Annual yield of apple fruit in a hilly district.
 - Study of nutrient contents of food consumed by the residents in a city.
- E4) Consider a population consisting of 5 villages, the areas (in hectares) of which are given below

Village	A	B	C	D	E
Area	760	343	657	550	480

Enumerate all possible WOR samples of size 3. Also write the values of the study variable (area) for the sampled units.

List all the WR samples of size 3 along with their area values.

Now, we consider the simplest of the random sampling procedures i.e., simple random sampling.

12.4 SIMPLE RANDOM SAMPLING

If each sample among the all possible samples has the same chance of being selected, then the associated method is called **simple random sampling**.

In simple random sampling procedure with unit by unit selection, every unit has got equal chance (probability) of selection at every draw. However, the converse is not true i.e. there are sampling schemes in which every unit gets the same chance of selection but they are not simple random sampling methods e.g. the systematic sampling. You shall learn about such sampling methods later in this unit. We now try to answer the question which may be occurring in your mind.

How to select a simple random sample (SRS)?

We consider the selection of a simple random sample through unit by unit selection method. At every draw equal probabilities are to be assigned to the available sampling units of the population. Thus, a pre-requisite for the selection is a random device by which selections are to be made. The most commonly used procedures for selecting a SRS are (1) lottery method, (2) through the use of random number tables. Let us discuss these methods one by one..

Lottery method

As the name suggests, units from the given frame (of size N) are selected using any procedure of generating a number randomly through lottery procedures. The simplest method may be writing down N numbers on identical slips of papers and drawing one of the slips after thoroughly mixing the slips. The number on the selected slip indicates the unit selected. For instance, suppose we have a population of 400 individuals and we wish to draw a random sample of 40 individuals. We can number the individuals of the population serially from 1 to 400. We can then take 400 identical slips of paper, write numbers 1 to 400 on them, put them in a box, mix them thoroughly and pick out 40 slips, one by one without looking. This gives us a random sample of 40 individuals. In with replacement procedure, the slip is replaced before the next draw while in case of without replacement it is not replaced. The sampling is continued till desired number of units are selected.

Any other randomization device such as pack of cards or random disc etc, may be used. However, the procedure becomes cumbersome if large number of selections is to be made as numbering of the slips become inconvenient and one has to be careful to see that the slips are thoroughly mixed after each draw. This method is not so common

in random selections. We now discuss another method which makes use of random number tables.

Through random number tables

Before discussing the method based on the use of random number tables you may like to know what random numbers are? Random numbers are numbers generated by a random procedure involving repeated independent trials. Such numbers are generated with the help of random digits 0 through 9. When we say random digits 0 through 9 it is assumed that a trial of the procedure yields each of the ten digits with probability 0.1. One simple way of generating these random digits is to take ten cards of the same size and write the digits 0 through 9 on the cards so that each card has a different digit. Then take a large hat, say, toss in the cards and mix them well. Now choose a card at random from the hat. Write down on a piece of paper the digit appearing on the card you have chosen. Put the card back into the hat and mix the cards again. Repeat the procedure by choosing a card at random, writing down the digit appearing on the card, replacing, mixing, choosing again, and so on. The string of digits we write down constitutes a string of random digits because it has been produced by a random device supposed to yield each digit with probability 0.1 in independent trials. Random digits can also be produced by using a modified roulette wheel in which the wheel is divided into ten equal parts, each one corresponding to one of the ten digits.

Given random digits, we can get more complicated random numbers. Suppose we have generated the sequence 3217900597 of ten digits. Then each digit is random, and also the two digit numbers 32, 17, 90, 05, 97 obtained by taking the numbers two at a time are random numbers because they have been produced by a random procedure so that each of one hundred two-digit numbers 00 through 99 has the probability 0.01 of appearing, and moreover selection of these two digit numbers are independent. Taking the original ten digit sequence and choosing the two digits at a time going backward to get 79, 50, 09, 71, 23 also gives random two-digit numbers. In the same way you might think of some other ways to get two-digit numbers using the generated string, as long as the method does not use the same selection more than once.

-
- E5) Give two more sequences of 5 two digit random numbers obtained by using the string 3217900597 of ten digits.
-

In a similar manner, random numbers with three, four or even more digits can be obtained by using the given string of random digits. There are several standard random number tables available which give the arrangement of these numbers in a rectangular manner. Some of these which are commonly used are prepared by Tippett (1927), Fisher and Yates (1938), Kendall and Smith (1939), Rand Corporation (1955), and Rao et al. (1974). One such random number tables is reproduced in Appendix A. You may note that this random number table can be used as single digit numbers or two digit numbers or, three or four digit numbers depending on the size of the population you are sampling from. We shall now illustrate the use of random number tables for selecting samples.

We shall discuss here three commonly used methods of using random number tables for selection of simple random samples.

Direct Approach. The first step in the method is to assign serial numbers 1 to N to the N population units. If the population size N is made up of K digits, then consider K digit random numbers, either row wise or column wise, in the random number table. The sample of required size is then selected by drawing, one by one, random numbers from 1 to N, and including the units bearing these serial numbers in the sample.

Problem 1: Consider a population of 56 households. Select a simple random sample of 10 households by with replacement as well as without replacement methods.

Solution: Here the sampling unit is a household. The first step is to serially arrange the households if they are not already arranged so. Since, the population size is a number consisting of two digits we have to use two digit random number table. Alternatively, in the table of random numbers (Appendix A) the first two digits of any column can be used randomly. While using these tables, it is advisable to take a blind start on the table by placing your finger on the table with closed eyes – let it be column 7, row 6 of the first page of Appendix A. Then the first number is 20 and going down the page subsequent random numbers between 1 to 56 are 12, 03 etc.

By selecting first 10 random numbers from 1 to 56, without discarding repetitions for **with replacement procedure (WR)**, we obtain the serial numbers of the households in the sample. These are given below:

20 12 03 16 30 15 24 37 01 15

For **without replacement procedure (WOR)**, repetitions have to be avoided and the number 15, when appears again at the tenth draw, should be dropped. The next two digits are then chosen. Thus, a WOR sample will consist of:

20 12 03 16 30 15 24 37 01 07

————— X —————

This procedure may involve number of rejections of random numbers, since zero and all the numbers greater than 56 appearing in the table are not considered for selection. The use of random numbers has, therefore, to be modified. We now discuss two of the commonly used modified procedures,

Remainder Approach. In this method if the population size N is a K digit number, then first we have to determine the highest K digit multiple of N . Let it be N' . Then a random number r is selected, such that $1 \leq r \leq N'$. This number r is then divided by N and let the remainder be R . The unit bearing the serial number equal to the remainder R , is then considered as selected. If remainder is zero, the last unit is selected. As an illustration, let $N=24$. Here N is a two digit number. The highest two digit multiple of 24 is 96. Let us now choose a number between 1 and 96 say, 83. On dividing 83 by 24, we get the remainder as 11. Therefore, the unit bearing serial number 11 is selected in the sample. Then another number between 1 and 96 is selected and the process is repeated till the sample of required size is selected. As before, the repeated selections of population units in the sample are permitted for WR sample, whereas they are rejected and only distinct units selected for a WOR sample. With a little variation in this method we consider another method.

Quotient Approach. As before, let N be a K digit number and N' be the highest K digit multiple of N , such that $N' = Nm$ for some integer m . Select a random number r from 0 to $N'-1$. Then the unit having serial number $(Q+1)$ is included in the sample, where Q is the quotient when r is divided by m . For instance, if $N=24$ then $N'=96$ and $m=4$. Let a random number $r=49$ be chosen from 0 to 95. Then dividing 49 by 4 one gets the quotient $Q=12$. The unit bearing serial number $(Q+1)=13$ is then selected in the sample. The process is repeated by selecting each time a new number r from 0 to 95 till a sample of required size is obtained.

As we have mentioned earlier while using the random number tables, any starting point can be used, and one can move in any predetermined direction along the rows or columns. However, normally as a convention, column-wise selection is followed. If

more than one sample is to be selected in any problem, each should have an independent starting point.

Besides the methods discussed above, some more methods for sample selection are available in the literature. However, being operationally inconvenient, these are usually not employed in practice.

In order to get conversant with the methods discussed above for selecting a simple random sample, you may try the following exercise. While doing this exercise you will also get convinced that the number of rejections in quotient and remainder approach are much less as compared to the direct approach.

-
- E6) Select a simple random sample of 10 households from the same population of 56 households by WR and WOR methods, using remainder approach and quotient approach.
-

The whole purpose of sampling is to collect information about the population from which the sample is drawn. It is used to study the unknown characteristics in the population called **parameters**. However, a sample cannot tell us about the population parameters exactly, it can only estimate parameters of the population. In the next section we shall see how these estimations are done.

12.5 ESTIMATION OF POPULATION PARAMETERS

In order to infer about population parameters, we compute various quantities from the sample. These computed quantities from the sample are called **statistics**. In general, we can say that a statistic is any quantity computed from a sample. Values such as mean, variance and standard deviations derived from samples are sample statistics which are then used to estimate population parameters and hence are called **estimators**.

Some of the important population parameters required to be estimated are population mean, variance and proportion. When the population is of size N , comprising of units with variate values Y_1, Y_2, \dots, Y_N , then **population mean** and **variance** are

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i = N^{-1}Y, \text{ where } Y = \text{Total population and } \sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (1)$$

respectively. The corresponding formulas for **sample mean** and **variance** are given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (2)$$

respectively for a sample of size n with variate values y_1, y_2, \dots, y_n . You may also note here that a **parameter is a fixed unknown quantity**. For example, the average height of the population of adult Indian males at a given time, has a single fixed value.

A **statistic** on the other hand is a **variable quantity**. The value of a statistic (or, an estimator) computed from different samples would differ from sample to sample. We know that the heights of adult Indian males vary. If two samples of the same size are drawn from this population then it may happen that one sample has a few more taller people than the other. Hence, the average height computed from one sample is likely to be different from that computed from the other. Thus, there is a need to study this

variability in the statistic if it is to be used as an estimator. In other words, we can say that there is a need to study the distribution known as sampling distribution of an estimator.

The sampling distribution of an estimator helps in defining certain desirable criteria for goodness of an estimator. One of the most important criteria is unbiasedness. The estimator is said to be unbiased for the parameter t , if $E(t) = t$. where $E(\cdot)$ stands for expectation. This expectation is computed by averaging the value of t over all possible samples. The criterion of unbiasedness ensures that on an average the estimator will take value equal to the unknown population parameter t . We now illustrate the concept of a sampling distribution through an example.

Problem 2 : Consider a simple random sample (WOR) of two households from a population of five households having monthly income (in rupees) as follows :

Household	1	2	3	4	5
Income (rupees)	1560	1490	1660	1640	1550

Enumerate all possible samples (WOR) of size 2 and show that the sample mean gives an unbiased estimate of population mean.

Solution: Population mean $\bar{Y} = \frac{1560 + 1490 + 1660 + 1640 + 1550}{5} = \frac{7900}{5} = 1580$

All possible samples and their corresponding sample means in this case are presented in Table-2.

Table 2 : All samples and their corresponding sample means in SRSWOR

(N=5, n=2)

Sample No.	Units in sample	Probability	Sample observations		Sample mean $\bar{y} = \frac{y_1+y_2}{2}$
			y_1	y_2	
1	1,2	1/10	1560	1490	1525
2	1,3	1/10	1560	1660	1610
3	1,4	1/10	1560	1640	1600
4	1,5	1/10	1560	1550	1555
5	2,3	1/10	1490	1660	1575
6	2,4	1/10	1490	1640	1565
7	2,5	1/10	1490	1550	1520
8	3,4	1/10	1660	1640	1650
9	3,5	1/10	1660	1550	1605
10	4,5	1/10	1640	1550	1595
Average			1580		

It may be seen that the average of sample means (1580) is equal to the population mean. This shows the unbiased nature of the sample mean as an estimator of population mean.

X

In the same way unbiasedness can be shown in the case when all possible samples of size 2 are drawn with replacement. We are leaving this for you to do it yourself.

-
- E7) In Problem 2 above, enumerate all possible samples (WR) of size 2 and show that the sample mean is an unbiased estimator of the population mean.
-

The **sampling variance** is the variance of the sampling distribution of the estimator. It measures the divergence of the estimator from its expected value. If $\hat{\theta}$ is an estimator of θ then,

$$V(\hat{\theta}) = E(\hat{\theta} - E\hat{\theta})^2$$

The positive square root of sampling variance is termed **standard error (SE)**.

$$SE(\hat{\theta}) = \sqrt{V(\hat{\theta})}$$

Thus, standard error is the standard deviation of the sampling distribution. It measures the precision of the estimator particularly in view of the fluctuations due to specific sampling design.

In case of **simple random sampling with replacement (SRSWR)**. **Variance of \bar{y}** can be written as

$$V(\bar{y}) = \frac{\sigma^2}{n} \text{ where, } \sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (3)$$

and **estimated variance** is

$$V(\bar{y}) = \frac{s^2}{n} \text{ where, } s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4)$$

For **simple random sampling without replacement (SRSWOR)**

variance of \bar{y} is

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N} \right) S^2 \text{ where, } S^2 = \frac{N\sigma^2}{N-1} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 \quad (5)$$

and **estimated variance** is

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N} \right) s^2 \text{ where, } s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (6)$$

To have a better understanding of the above formulas, you may try the following exercises.

- E8) Suppose we have a population of 5 students enrolled for statistics course and a counsellor wants to find the average amount of time spent by each student in preparing for classes each week. The amount of time (in hours) each student spends per week is given by 7,3,6,10 and 4. If the counsellor takes a sample of three students WOR, obtain the sampling distribution of the sample mean. Compute the population mean and the mean and standard error of the sampling distribution.
- E 9) In the data of Example 2, show that the sample mean square (s^2) is an unbiased estimator of the population mean square (S^2).

Estimation Of Proportion

Sometimes interest lies in estimating the population proportion. Examples, such as, proportion of persons below poverty line or proportion of female members in a particular group or proportion of persons getting degrees through distance education etc. are very common. In all these examples the population is considered as divided in two parts on the basis of an attribute.

For instance, a crop field may be irrigated or not irrigated. If it is irrigated, we say that it possesses the characteristic 'irrigation'. If it is not irrigated, we say that it does not possess the particular characteristic of irrigation. If we are interested in estimating the proportion of irrigated fields, the population of N fields can be defined with variate y_i as

$$y_i = 1, \text{ if the field is irrigated}$$

$$= 0, \text{ otherwise}$$

If the total number of irrigated fields be N_1 out of N then

$$\sum_{i=1}^N y_i = N_1$$

$$\text{Thus, } \bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i = \frac{N_1}{N} = P = \text{proportion of irrigated fields.}$$

Thus, the problem of estimating a population proportion becomes that of estimating a population mean by defining the variate as above. Now if a simple random sample of size n is taken from the population and if n_1 units out of n possess that characteristic, then sample proportion is given by

$$p = \frac{n_1}{n}$$

$$\text{Thus, } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{n_1}{n} = p.$$

It follows that, in SRSWR as well as SRSWOR, p is an unbiased estimator of P . The variances and estimator of variances in case of WR and WOR procedures are given as follows.

For SRSWR

Variance of p is

$$V(p) = \frac{PQ}{n}, \text{ where } Q = 1 - P, \quad (7)$$

and estimated variance is

$$V(p) = \frac{pq}{n-1}, q = 1 - p \quad (8)$$

For SRSWOR

Variance of p is

$$V(p) = \frac{N-n}{N-1} \left(\frac{PQ}{n} \right), Q = 1 - P \quad (9)$$

and estimated variance is

$$v(p) = \frac{N-n}{N} \left(\frac{pq}{n-1} \right), q = 1 - p \quad (10)$$

We now take up an example to illustrate what we have discussed above.

Problem 3: Obtain the sampling distribution of the sample proportion and the standard error of the proportion of households having monthly income more than Rs.1550 in the population given in Problem 2 by considering simple random sample (WOR) of two households.

Solution: In the population of 5 households the income of household 1,3 and 4 is more than Rs.1550. Population proportion of interest to us is $P = 3/5$.

Let us now work out the sampling distribution of the sample proportion by getting the sample proportion of households with income exceeding Rs.1550 from each of 10 possible samples listed in Table 2. In each sample we score a household as 0 if its income is Rs.1550 or less and as 1 if it exceeds Rs.1550. Then the mean score in each sample shown in Table-3 below gives the sample proportions.

Table-3: Computation of the sample proportions

Sample No.	Score	Mean Score
1	1,0	½
2	1,1	1
3	1,1	1
4	1,0	½
5	0,1	½
6	0,1	½
7	0,0	0
8	1,1	1
9	1,0	½
10	1,0	½

Standard error $SE(p)$ of the proportion given by Formula (9) is

$$\begin{aligned} SE(p) &= \sqrt{\frac{N-n}{N-1} \frac{P(1-P)}{n}} \\ &= \sqrt{\frac{5-2}{5-1} \left(\frac{3}{5}\right)\left(1-\frac{3}{5}\right)\frac{1}{2}} \\ &= \sqrt{\frac{9}{100}} = \frac{3}{10} = 0.3 \end{aligned}$$

X

Why don't you try this exercise now

- E10) IT facility committee of IGNOU has a total of eight members whose ages in years are 27, 32, 33, 26, 43, 52, 28 and 25. The committee has a rule which requires a minimum age of 33 for a member to be the chairperson. Assume that a simple random sample of size 4 is selected to provide an estimate of the population proportion eligible to be chairperson. Find the mean and standard deviation of the sampling distribution.
-

Sample size determination

While planning a sample survey, a decision has to be made regarding the sample size in the initial stages. Having decided about the method of selection, one has to determine about the sample size in view of the resources available as well as the desired level of precision of the estimators. Larger sample sizes will involve more cost in data collection as well as data analysis while smaller sample sizes will reduce the precision of estimate. Therefore, a balance has to be struck between cost and precision, while deciding about the sample sizes. We consider a simple example to explain the idea behind principles involved in determination of sample size.

An anthropologist wants to study the inhabitants of some island. He wishes to estimate the percentage of inhabitants belonging to blood group O. A simple random sample is to be selected. How large should the sample be?

Some related questions are pertinent here. How accurately does the anthropologist wish to know the percentage of people with blood group O? Suppose, he answers that he will be contented if the percentage is correct within a tolerance limit d of $\pm 5\%$. It is also to be understood that even with this specification of tolerable limit of error it is not possible to ensure that the estimates are obtained in this margin in 100% of cases. A level of confidence has therefore to be attached with the estimates. Let confidence level $1 - \alpha$ be 95% associated with the estimates.

Let us assume that p is normally distributed about P . It will then lie in the range $(P \pm 2\sigma_p)$ where σ_p is the standard deviation of p , apart from a one in twenty chance (i.e. apart from a probability of α).

In case of **SRSWR**, $\sigma_p = \sqrt{\frac{PQ}{n}}$. Hence we can put

$$2\sqrt{\frac{PQ}{n}} = \frac{5}{100} \text{ or } n = \frac{4PQ}{25} \times 100 \times 100.$$

At this stage some idea about P is needed in order to determine the sample size n . Fortunately, we do not need very accurate estimate of P for this purpose. In fact, if P

lies between 0.3 to 0.6 then, the determined sample size lies between 336 to 400. To be on the safe side, 400 may be taken as the initial estimate of n .

Note that the maximum value of PQ with $0 \leq P \leq 1$, $Q = 1 - P$ is attained at $P = 0.5$ and its value is 0.25.

In case of SRSWOR, for estimation of P , the formula for sample size is given as

$$n = \frac{t^2 PQ/d^2}{1 + \frac{1}{N} \left(\frac{t^2 PQ}{d^2} - 1 \right)}$$

where d is the margin of tolerable error and t is the abscissa of the normal curve that cuts off an area of α at the tails, $(1-\alpha)$ being the confidence level of the estimate. If N is large, a first approximation to n is $n_0 = \frac{t^2 P Q}{d^2}$.

In the example considered above tolerable error is within 5% so $d = 0.05$. We want this at the level of confidence of 95%, or in other words $\alpha=0.05$. Assume that $P = 0.5$. From the standard normal distribution, the value of the variate corresponding to the two-sides tail of 5% is 1.96 or approximately 2 and hence

$$t=2 \text{ Thus, } n_0 = 4 \times \frac{0.5 \times 0.5}{(0.05)} = 4 \times 100 = 400.$$

In simple random sampling, units are selected randomly at each draw. Now we shall discuss a sampling technique which has a nice feature of selecting the whole sample with just one random start.

12.6 SYSTEMATIC SAMPLING

In simple random sampling units are drawn randomly at every draw. In many situations, it may be desirable to select a sample in a systematic way. For example, if we want to have an even spread in terms of spatial distribution, a systematic selection may ensure that units maintain a uniform distance between selected units. In **systematic sampling**, one unit is selected randomly and subsequent units are selected according to a pre-determined system. Invariably uniform distance is adopted for pre-assigned system. Systematic samples actually provide an improvement over simple random samples as the samples are spread more evenly over the entire populations. We now discuss sample selection procedures for systematic sampling.

12.6.1 Linear Systematic Sampling

The most commonly adopted procedure of systematic sampling is **linear systematic sampling**. We shall explain the method through an example.

Problem 4 : Consider a population of 12 households from which a sample of 3 households is to be selected.

Solution: Let the households be arranged serially from 1 to 12. These households are now rearranged in 3 rows of 4 columns as follows:

1	2	3	4
5	6	7	8
9	10	11	12

Then, for selecting a systematic sample of size 3, we select a random number r (say) between 1 to 4. Starting with r , every 4th unit is selected. Thus, if $r=3$, then the units selected are 3, 7(=3+4), and 11(=7+4). Thus, if r is selected the entire column of units consisting of r is selected.

— X —

In general, this method is applicable if the population size N is a multiple of the sample size n i.e. $N = nk$ where k is an integer. The random number r is selected between 1 to k . Here, r is called a **random start** and k is called **sampling interval**. The sample then comprises of the units $r, r+k, r+2k + \dots + r + (n-1)k$. The technique will generate k systematic samples with equal probability. The method is known as linear systematic sampling as N units are assumed to be arranged sequentially on a line.

The method is specially suitable in forestry where for estimating the volume of timber this method is used for selection of area units. Some other applications are in industries where items for sample checks are selected systematically in a production process. The concept of systematic sampling is not only confined to spatial distributions. It can also be done over time. In fact in one of the applications in estimation of fish catch from marine resources, sampling of boats on landing centres is carried out systematically over time. Boats arriving every two hourly on selected landing centres are observed.

In the method described above, if N is not a multiple of n , then it may not be possible to get samples of equal size. For example, if $N = 14$ and $n = 3$ then the method described above would lead to following arrangement of units in a $n \times k$ table as follows:

1	2	3	4
5	6	7	8
9	10	11	12
13 14			

In this case if randomly selected number r between 1 to 4 is 1 then the sample is 1,5,9,13 while if $r=3$, then the sample is 3,7,11. Thus, samples are not of the same size. Sample size is either 4 or 3 depending on the value of r . As an improvement to this method we shall discuss circular systematic sampling

12.6.2 Circular Systematic Sampling

To overcome the difficult of varying sample size in a situation when $N \neq nk$ the procedure is modified slightly by which a sample of constant size is always obtained. This procedure is known as circular systematic sampling

In this method, the N units may be regarded as arranged round a circle. A random start is taken between 1 to N and thereafter every k^{th} unit, k being an integer nearest to $\frac{N}{n}$, in a circular manner is selected until a sample of n units is chosen. Suppose that a unit with random number i is selected. The sample will then consists of the units corresponding to the serial numbers.

$$i + jk; \text{ if } i + jk \leq N$$

$$i + jk - N; \quad \text{if } i + jk > N \quad \text{for } j = 0, 1, \dots, (n-1)$$

This method is applicable, even if $N \neq nk$. To illustrate the method, we consider the following example:

Problem 5: Consider a population of 14 households from which a sample of size 5 is to be selected.

Solution: Here $N=14$, $n=5$, $k=3$. Consider Fig.1:

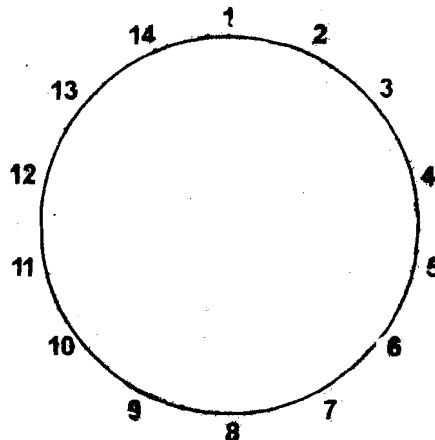


Fig.1

Let the random start be 7. Then the selected sample is 7, 10, 13, 2, 5. If we start from 9 then the selected sample is 9, 12, 1, 4, 7. Like this we can have 12 more samples as the total number of possible samples in this case is $N = 14$.

— X —

The above method has got the advantage of providing samples of the given size irrespective of the random start. In case of linear systematic sampling the number of possible samples is k while in case of circular systematic sampling it is N . When N is a multiple of n then linear systematic sampling is normally preferred although one could also go for circular systematic sampling. However, when N is not a multiple of n then one should necessarily go for circular systematic sampling.

12.6.3 Advantages and Limitations of Systematic Sampling

The systematic sampling has the nice feature of operational convenience because the selection of the first unit determines the whole sample. This operation is easier to understand and can be speedily executed in relation to simple random sampling. Secondly, systematic samples are well spread over the population and there is no risk that any large part of the population will be left unrepresented. For populations with linear trend, systematic sampling is more efficient in comparison to simple random sampling.

Systematic sampling should, however, be cautiously used in case the population exhibits a periodic trend. For periodic populations, the efficiency of the systematic sampling depends upon the value of the sampling interval. If the sampling interval coincides with the period, the sample will contain identical units and consequently the systematic sampling performance becomes very poor. If, however, the sampling interval is an odd multiple of half the period, the systematic sampling becomes most effective.

A serious limitation of this scheme lies in its use with populations having unforeseen periodicity, which may substantially contribute to the bias in the estimate of mean/total. Another serious limitation of the sampling scheme, as mentioned earlier, is that the variance of the estimator cannot be estimated unbiasedly.

You may now try this exercise.

-
- E11) A sample of size 4 is to be selected from a population of 11 households. List all the possible sample by (i) linear systematic sampling ii) circular systematic sampling
-

We now end this unit by giving a summary of what we have covered in it.

12.7 SUMMARY

In this unit, we have learnt

- 1) The preliminary concepts and definitions for simple random sampling
- 2) Method of selecting a simple random sample
- 3) How to estimate the population mean/total
- 4) The method of estimating population proportions
- 5) How to determine sample size in case of simple random sampling
- 6) The basic concept of systematic sampling
- 7) How to select a linear systematic sample
- 8) The method of selecting a circular systematic sample
- 9) The advantages and limitations of systematic sampling

12.8 SOLUTIONS/ANSWERS

- E1). i) Collection of all the households in a locality/city is the population and the individuals are the households
- ii) Total number of crates loaded in a truck is the population and the individuals are the crates.
- iii) All the bulbs produced by the company during a shift is the population and the individuals are the bulbs.
- E2) Advantages: less expensive in terms of time, money and energy, less cumbersome, free of nonsampling errors. Also in case of destructive testing, sampling is the only method.
Disadvantages: may suffer from sampling errors. The estimate is only an approximation to the true value.
- E3) a) The collection of all trees in the forest is the population, a tree is the individual sampling unit and a list of all trees is the sampling frame.
b) Total number of apple trees in a district is the population, an apple tree is the sampling unit and list of trees is the sampling frame.

Sampling

c) Collection of all the household in a city consuming the food is the population, an individual household is the sampling unit and a list of all households selected for a sample is the sampling frame.

- E4) a) Table 3: All samples and their corresponding sample means in SRSWOR ($N=5$, $n=3$) (order of sample units is ignored)

Sample No.	Units in Sample	Sample areas			$\bar{y} = \frac{y_1 + y_2 + y_3}{3}$
		y_1	y_2	y_3	
1	A,B,C	760	343	657	586.67
2	A,B,D	760	343	550	551
3	A,B,E	760	343	480	527.67
4	B,C,D	343	657	550	516.67
5	B,C,E	343	657	480	493.33
6	C,D,E	657	550	480	562.33
7	C,D,A	657	550	760	655.67
8	D,E,A	550	480	760	596.67
9	A,C,E	760	657	480	632.33
10	B,D,E	343	550	480	457.67
	Average				558

- b) Similarly make a table for all samples of size 3 with replacement. There will be 5^3 samples in all.

- E5) One sequence could be 21,79,00,59,73. Similarly give another.

- E6) **Remainder approach:** For $N=56$, the highest two digit multiple of N is N itself. Using Appendix A select a two digit random number r , s.t. $1 \leq r \leq 56$.

$r=44$ is one possibility. Also $r/N = \frac{44}{56}$ gives quotient as 0 and remainder is 44. Thus select the unit with serial number 44. Likewise you can select other 9 units also. One such simple random sample (WOR) of 10 households selected could be a sample of households with serial numbers

44, 49, 40, 15, 12, 38, 29, 52, 22, 50

and a sample of 10 households (WR) could be

44, 49, 40, 15, 44, 12, 38, 29, 52, 22

Quotient approach: In this case $m=1$. So if a random number r is selected s.t $1 \leq r \leq 56$, say $r=44$, then dividing r by m we get $Q = 44$ and selected unit is the unit with serial number $44+1 = 45$. Likewise you can select a SRS of 10 households (WR) and (WOR) from the population of 56 households.

- E7) Average of sample means (1580) is equal to the population mean and thus the sample mean is an unbiased estimator of the population mean. Calculations are shown in Table 4 in the next page.

Table 4: All samples and their corresponding sample means in SRSWR (N=5, n=2)
(order of sample units ignored)

Sample No.	Units in Sample	Probability	Sample observations		Sample mean
			y ₁	y ₂	
1	1,2	1/10	1560	1490	1525
2	1,3	1/10	1560	1660	1610
3	1,4	1/10	1560	1640	1600
4	1,5	1/10	1560	1550	1555
5	2,3	1/10	1490	1660	1575
6	2,4	1/10	1490	1640	1565
7	2,5	1/10	1490	1550	1520
8	3,4	1/10	1660	1640	1650
9	3,5	1/10	1660	1550	1605
10	4,5	1/10	1640	1550	1595
11	1,1	1/10	1560	1560	1560
12	2,2	1/10	1490	1490	1490
13	3,3	1/10	1660	1660	1660
14	4,4	1/10	1640	1640	1640
15	5,5	1/10	1550	1550	1550
Average					1580

E8) Distribution of Sample mean in sample size 3

Sample mean value	No. of Samples giving this means (frequency)	Relative frequency
4.33	1	1/10
4.67	1	1/10
5.33	1	1/10
5.67	2	2/10
6.33	1	1/10
6.67	2	2/10
7.0	1	1/10
7.67	1	1/10
Total	10	1

Sampling

$$\text{Population mean } \mu = \frac{7+3+6+10+4}{5} = \frac{30}{5} = 6.$$

The mean of the above distribution

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum_{i=1}^5 (y_i - \mu)^2}{N}} = \sqrt{\frac{(7-6)^2 + (3-6)^2 + (6-6)^2 + (10-6)^2 + (4-6)^2}{5}} \\ &= 2.45 \quad (N=5, n=3) \\ S.E &= \frac{2.45}{\sqrt{3}} \sqrt{\frac{5-3}{5-1}} = 1.00 \text{ (approximately)} \\ &= \frac{4.33 \times 1 + 4.67 \times 1 + 5.33 \times 1 + 5.67 \times 2 + 6.33 \times 1 + 6.67 \times 2 + 7.0 \times 1 + 7.67 \times 1}{10} \\ &= \frac{60}{10} = 6\end{aligned}$$

$$S.E = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (\text{Using formula (5)})$$

where σ is the population standard deviation

$$\begin{aligned}E9) \quad \text{Sample Variance } s^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= 2450 + 5000 + 3200 + 50 + 14450 + 11250 + 1800 + 200 + 6050 + 4050 \\ &= 48500\end{aligned}$$

average of 10 sample mean squares gives $E(s^2)$.

$$\text{Here } E(s^2) = 48500/10 = 4850 \quad (i)$$

Also, population mean square =

$$\begin{aligned}S^2 &= \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{4} (400 + 8100 + 6400 + 3600 + 900) \\ &= \frac{1}{4} \times 19400 = 4850 \quad (ii)\end{aligned}$$

Thus from (i) and (ii)

$$E(s^2) = S^2$$

i.e. sample mean square provides an unbiased estimator of the population mean square.

E10) The mean of the Sampling distribution is the population proportion P ,

$$P = \frac{3}{8}$$

S.E for the sampling distribution

$$= \sqrt{\frac{P(1-P)}{n}} \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{\frac{3}{8}(1-3/8)}{4}} \sqrt{\frac{8-4}{8-1}} = \sqrt{\frac{3}{8} \times \frac{5}{8} \times \frac{1}{4}} \sqrt{\frac{4}{7}} = 0.183$$

- E11) i) $N=11, n=4, k = \frac{11}{4} = 3$ (approx). Arranging the units in 4 rows of 3 columns each (except for the last row) we get table as follows:

1	2	3
4	5	6
7	8	9
10	11	

Selecting a number r between 1 and 3, possible samples are 1,4,7,10; 2,5,8,11; 3,6,9 of size 4 or 3.

- ii) Here $N=11, n=4, k=3$. Consider Fig.2

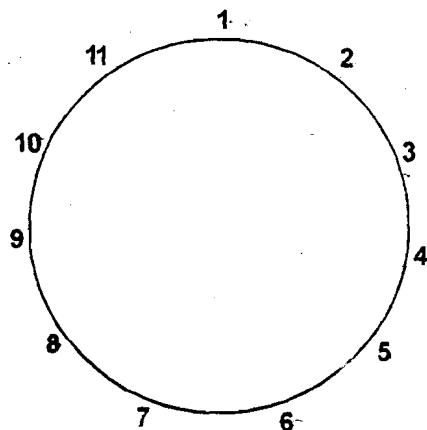


Fig.2

Let the random start be 2. Then sample selected is 2,5,7,10. If random start be 5 then sample selected is 5,8,11,3. Likewise you can write the remaining 9 samples as the total number of possible samples is $N=11$.

Appendix A : Random Numbers

Column Numbers									
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
3436	6833	5809	9169	5081	5655	6567	8793	6830	1332
6133	4454	2675	3558	7624	5736	2184	4557	0496	8547
9853	3890	5535	3045	9830	5455	8218	9090	7266	4784
5807	5692	6971	662	6751	5001	5533	2386	0004	2855
6291	0924	1298	7386	5856	2167	8299	9314	0333	8803
4725	9516	8555	0379	7746	9647	2010	0979	7115	6653
7697	6486	3720	6191	3552	1081	6141	7613	5455	3731
3497	2271	9641	0304	4425	6776	1205	2953	5669	1056
8940	4765	1641	0606	4970	7582	7991	6480	2946	5190
1122	6364	5264	1267	4027	4749	0338	8406	1213	5355
4333	0625	3947	1373	6372	9036	7046	4325	3491	8989
7685	1550	0853	4276	1572	9348	6893	2113	8285	9195
0592	8341	4430	0496	9613	2643	6442	0870	5449	8560
3506	0774	0447	7461	4459	0866	1698	0184	4975	5447
8368	2507	3565	4243	6667	8324	3063	8809	4248	1190
2630	1112	6680	4863	6813	4149	8325	2271	1963	9569
3883	3897	1848	8150	8184	1133	6088	3641	6785	0658
1123	3943	5248	0635	9265	4052	1509	1280	0953	9107
1167	9827	4101	4496	1254	6814	2479	5924	5071	1244
7831	0877	3806	9734	3801	1651	7169	3974	1725	9709
2487	9756	9886	6776	9426	0820	3741	5427	5293	3223
1245	3875	9816	8400	2938	2530	0158	5267	4639	5428
5309	4806	3176	8397	5758	2503	1567	5740	2577	8899
7109	0702	4179	0438	5234	9480	9777	2858	4391	0979
8716	7177	3386	7643	6555	8665	0768	4409	3647	9286
9499	5280	5150	2724	6482	6362	1566	2469	9704	8165
3125	4552	6044	0222	7520	1521	8205	0599	5167	1654
3788	6257	0632	0693	2263	5290	0511	0229	5951	6808
2242	2143	8724	1212	9485	3985	7280	0130	7791	6272
0900	4364	6429	8573	9904	2269	6405	9459	3088	6903
7909	4528	8772	1876	2113	4781	8678	4873	2061	1835
0379	2073	2680	8258	6275	7149	6858	4578	5932	9582
0780	6661	0277	0998	0432	8941	8946	9784	6693	2491

8478	8093	6990	2417	0290	5771	1304	3306	8825	5937
2519	7869	9035	4282	0307	7516	2340	1190	8440	6551
2472	0823	6188	3303	0490	9486	2896	0821	5999	3697
8418	5411	9245	0857	3059	6689	6523	8386	6674	7081
8293	5709	4120	5530	8864	0511	5593	1633	4788	1001
9260	1416	2171	0525	6016	9430	2828	6877	2570	4049
6568	1568	4160	0429	3488	3741	3311	3733	7882	6985
6694	5994	7517	1339	6812	4139	6938	8098	6140	2013
2273	6882	2673	6903	4044	3064	6738	7554	7734	7899
6364	5762	0322	2592	3452	9002	0264	6009	1311	5873
6696	1759	0563	8104	5055	4078	2516	1631	5859	1331
3431	2522	2206	3938	7860	1886	1229	7734	3283	8487
4842	3765	3484	2337	0587	9885	8568	3162	3028	7091
8295	9315	5892	6981	4141	1606	1411	3196	9428	3300
4925	4677	8547	5258	7274	2471	4559	6581	8232	7405
5439	0994	3794	8444	1043	4629	5975	3340	3793	6060
2031	0283	3320	1595	7953	2695	0399	9793	6114	2091

Source: Rao, CR, Mitra, S.K., Maitthai, A. and Ramamurthy, K.G. (1974). Formulae and Tables for Statistical Work. Statistical Publishing Society. Indian Statistical Institute Calcutta.

UNIT 13 STRATIFIED SAMPLING

Structure	Page No.
13.1 Introduction	28
Objectives	
13.2 Stratified Sampling	29
Preliminaries	
Advantages	
13.3 Estimation of population parameters	32
13.4 Allocation of sample size	36
13.5 Construction of strata	38
13.6 Post-Stratification	40
13.7 Summary	41
13.8 Answers/Solutions	42

13.1 INTRODUCTION

As seen in the previous unit, *sampling* is a process by which we choose a *representative sample* from a given *population*. The main *objective of sampling* is to make an optimal use of the budget and other resources for a study to obtain as precise an estimate of a population parameter as possible.

On the basis of what you read in Unit 12, it is plausible to accept that if the *target population* is homogeneous with respect to the study variable, then the sample size requirement is likely to be small. However, as we know, a population is always associated with certain amount of variability. And so, if the population is heterogeneous, a *larger sample* is required to increase the precision of the population estimates. But, as you will agree, it is not possible always to adopt a *sampling design* needing *large samples*.

Many modifications have evolved from the central concept of *simple random sampling* that permit more precise inferences to be attained for different types of populations, especially wherein the *variability* within *small subpopulations* is small. One of the most practically useful designs is **stratified sampling**, in which we first divide the *target population* into *homogenous segments* and then, following any method of sample selection, choose samples from each of these *segments* (or *subpopulations*) independently in such a way that the total number of units pooled over all the selected groups is the desired sample size. The *small groups* thus formed are called **strata** and the process of forming strata, is called **stratification**.

In Sec.13.2, we shall discuss *preliminaries about stratification, broad principles adopted while using stratified sampling*, and some *advantages of stratified random sampling*. In Sec.13.3, we shall illustrate the use of certain relations commonly employed in the estimation of the population characteristics. In Sec.13.4, we shall discuss *equal allocation, proportional allocation, and optimum allocation*, which are some of the commonly used *methods of sample size allocation*. In Sec.13.5, some aspects consideration involved in the construction of strata are discussed. Finally, the unit is concluded with a discussion on the principle of *post-stratification*.

Objectives

After reading this unit, you should be able to

- discuss the basics, principles and advantages of stratified sampling;
- estimate the population mean, population total, and proportion alongwith their efficiency in stratified random sampling;

- allocate the total sample size using various methods of allocation;
- construct strata using a simple procedure (*Dalenius-Hedges rule*);
- discuss the principle of post-stratification.

13.2 STRATIFIED SAMPLING

As said in the introduction of the unit, *stratification is the process of partitioning the entire population into small groups, called strata, each containing units homogenous with reference to study variables under consideration*. That is, homogeneity within a stratum is based on the characteristic under study.

Quite often, strata are available in natural forms. For example, in Agricultural Surveys, geographically contiguous units form a stratum under the assumption that nearby units are likely to be homogenous due to similar agro-climatic conditions as well as similar cultivation practices.

However, in other situations strata are formed on the basis of related variables. Listed below are some of the practical situations where the use of stratified sampling is a common practice.

- a) In crop estimation surveys, geographically contiguous areas such as tehsils/talukas or groups thereof are taken as strata.
- b) In many socio-economic surveys, (within villages) small, medium, and large cultivators are taken as strata.
- c) In National Sample Surveys, which are conducted continuously with multi-subject surveys being conducted in successive rounds, strata are formed by grouping contiguous tehsils which are homogeneous with respect to population density, altitude above sea level and cultivation of food crops.
- d) A wide variety of maps from Indian National Atlas showing population density, food crops etc., are used for stratifying the population.
- e) In many other situations, the geographical and topographical considerations are taken into account for resorting to stratification.

We shall use the following definition for our discussion in this unit.

Definition. *The procedure of partitioning a given population into homogeneous groups, called strata, and then selecting samples independently from each stratum is known as stratified sampling. If a sample from each stratum is selected by random sampling, the procedure will be called stratified random sampling.*

The following is the list of some of the broad principles that one has to keep in mind while adopting stratified sampling.

1. For obvious reasons, the strata should be non-overlapping and should together comprise the whole population.
2. To minimise variance within a stratum, the units forming any stratum should be similar with respect to the study variable.
3. Sometimes administrative convenience may be considered as the basis for stratification. However, if such strata are not necessarily homogeneous, sub-stratification within each geographical stratum may be adopted based on homogeneity criteria using some ancillary information.

Try the following exercise now.

-
- E1) With the help of two examples, explain the concept of stratified sampling clearly stating the principles adopted in the process.
-

The **stratified sampling**, with all its advantages of convenience, flexibility, efficiency with respect to sampling variance as well as cost, has become an essential component in all sample surveys of practical importance.

13.2.1 Preliminaries

Practical considerations like *cost, administrative convenience, simplicity of the methods*, etc., are kept in mind while stratifying a population.

The essence of stratification is that it capitalizes on the known homogeneity of the *subpopulations*, so that only relatively small samples are required to estimate the characteristic for each subpopulation. These individual estimates are then combined to produce an estimate for the entire population.

To illustrate this point, let us consider the case of a city in which the northern districts are predominantly *low-income* areas and the southern districts are primarily *high-income* areas. So, to estimate the *average income* for the whole city, it is intuitively apparent that *relatively small simple random samples taken separately from the northern and southern districts* are likely to provide more accurate information than a single random sample taken from the entire city.

Also, it is clear from above that in situations when *variability within population is wide ranging* more precise inferences can be made using stratified sampling rather by using simple random sampling.

In general, the following four basic questions are important in the process of stratified sampling.

- a) *How to form the strata?*
- b) *How many strata to be formed?*
- c) *How to select the samples in each stratum?*
- d) *How to allocate the sample size to different strata?*

Of course, these fundamental questions are addressed with a purpose to minimise the sampling variance.

In the following illustration, we shall analyse a practical situation in order to understand the *importance* of above listed four questions in the context of stratified sampling.

Example 1: The case we are discussing is that of a *transport company*, which we shall refer to as A&B in our subsequent discussion. Common understanding is that if a shipment travels over several roads, the total freight charge is divided among all the transport companies sharing the responsibility of shipment. Also, it is acceptable that the computations involved in determining each transport company's revenue are cumbersome and expensive.

Hence, A&B decided to conduct a study to determine if the division of total revenue among several companies could be made accurately on the basis of a *sample survey* and at a substantial savings in clerical exercises. So, the purpose of such a study was to determine *how much of this total revenue belongs to A&B*.

A *waybill*, which is a document issued with every shipment of freight, gives details about the goods, route, and charges.

In one of the experiments during a *six-month* period, A&B studied the division of revenue for all shipments travelling over more than twenty districts. *The waybills, from which the amounts due each transport company can be computed, of these shipments constituted the population under examination.* The total number of waybills in the population, as well as the total freight revenue accounted for by the population of waybills, was known.

For the six-month period under study, there were nearly 23,000 waybills in the population. The amounts of the freight charges on these waybills vary greatly (some freight charges were as low as Rs.200 and others as high as Rs.2000) and, so, it was decided to follow the stratified sampling procedure.

Since the amount due the A&B on a waybill tended to be larger when the total amount on that waybill was large, the strata in this case were set up according to the amount of the total freight charge. Specifically, the strata formed were as given in Table 1 below.

Table 1. Formation of Strata.

Stratum	Waybills with freight charges (in Rs.)
1	0 to 200
2	201 to 400
3	401 to 700
4	701 to 1400
5	over 1401

Table 2. Proportion to be sampled.

Stratum	Proportion to be sampled (in %)
1	1
2	10
3	20
4	50
5	100

Stratified Sampling

The next problem before A&B was to decide *how large a sample from a stratum must be selected* so that the amount of the revenue due them could be estimated with a required amount of precision from as small a sample as possible. One piece of information needed for this task was the number of waybills in each stratum. The final *sample size allocation* decided on for the starta were as given in Table 2 above. As is clear from this table, more sample units are taken from the strata containing wider ranges of freight charges and smaller number of sample units are taken from the strata containing narrow ranges of freight charges.

To understand the method of sample size allocation adopted by A&B, consider Stratum-1. Here, the stratum contains waybills with charges less than Rs.200. As the variation between the waybills amounts is small, so, a small sample will provide adequate information about the amounts of all of the waybills in this stratum. On the other hand, Stratum-4, containing waybills with charges between Rs.701 and Rs.1400, has much greater variation. And, hence a larger sample is required from Stratum-4 to obtain adequate information about the amounts of all waybills.

Once the sample sizes allocation to different strata is determined, the next problem with A&B was to *select the samples from each stratum* i.e., construction of strata. At this stage, it is important to select the samples according to a procedure which facilitate the evaluation of the *sample statistics* as precisely as possible. That is, we should be able to judge how close the sample results are to the relevant population characteristics. Simple random sampling, you read in the previous unit, is one such procedure that can be applied to select samples from each stratum.

The A&B actually used a slightly different method of selecting waybills from each stratum, called *serial number sampling*. In this procedure, the sample from each stratum were selected according to certain *digits in the serial numbers* of the waybill. Since the serial numbers appear prominantly on the waybills, so, in comparison to other methods, this procedure for selecting the sample was found simple.

Try the following exercise.

- E2) Why A&B choose to go ahead with a stratified sampling? Also, give reasons why each stratum is relatively homogenous with respect to the amount of freight charges due the A&B company.

More generally, as is clear from above discussion, there could be certain population related *constraints* because of which we are led to consider the stratified sampling. And then, in the next step, the population characteristics guide us in *strata formation process*. Finally, on the basis of our practice with some of the basic sampling methods, we adopt an appropriate *procedure* for selecting samples from each stratum.

Once through with these three stages, we next make sample statistics and check the consistency of results with population estimates. About *sample statistics* and *population estimates* we shall talk in the next section.

Let us wind up this section with a discussion on some of the *advantages* of the stratified sampling.

We assume that each stratum contains waybills with total freight charges of roughly the same order of magnitude.

You will read about the three methods of *sample size allocation* in Sec.13.4

You will read about the method of *construction of strata* in Sec.13.5.

13.2.2 Advantages

Some of the advantages of stratified sampling are briefly described in the following list.

1. It is clear that exclusion of a proportion of the population under study may lead to a wrong estimation of the population parameters. And, as we divide the population into various strata and then draw samples from each stratum so formed, there is very little possibility that a part of the population remains completely excluded. Hence, *stratification* ensures that a better cross section of the population is represented in a sampling design. Certainly, this is not the case with any *unstratified sampling* procedure.
2. It is desirable in some situations to use different sampling designs in different strata. *Stratification* allows us to do so, thereby, enabling effective utilization of the available auxiliary information. It is particularly true, when the extent and nature of the available information vary from stratum to stratum. A separate estimates obtained for different strata are combined to get a precise estimate for the whole population.
3. By using a proper strata formation procedure, the variability within strata can be considerably reduced. So, the stratification normally provides more efficient estimates than any unstratified sampling. For example, when there are several extreme values for the study variable in a population, they are grouped into a separate stratum thereby reducing the variability within strata.
4. In case of stratified sampling, the cost of conducting the survey is expected to be less. This is particularly true when strata are formed keeping administrative convenience in mind. This facilitates the supervision and organisation of field work involved in the sampling process.
5. There may be different types of sampling problems in plains, deserts, and hilly areas. These may need different approaches for their resolution. Hence, it would be advantageous to form separate stratum for each of these areas.

Try the following exercise.

- E3) Discuss the above stated four advantages of stratified sampling, giving one example in each case.

In this section, we have discussed certain preliminaries related to the process of stratified sampling. Also, we talked about certain advantages of stratified sampling that it has over other sampling methods. Let us now discuss the method of *estimation of population parameters* in the next section.

13.3 ESTIMATION OF POPULATION PARAMETERS

As said above, stratified sampling has got the flexibility that *any method of sampling* can be used independently within a stratum. However, in this unit, we consider only those situations wherein *simple random sampling without replacement (SRS-wor)* is used for selecting a sample from each stratum.

Let a (finite) population (under study) contain N units and suppose this population is divided into L number of strata (by any method), each containing units homogenous with respect to certain characteristics in question. Also, in the following table, the suffix h stands for the h^{th} stratum ($h = 1, 2, \dots, L$) and the suffix i will indicate the i^{th} unit within a stratum. The *notations* defined in Table 3 are fixed for our convenience and future use.

Since samples in stratified sampling are drawn independently from each stratum, the estimates of strata *means*, *totals* and *proportions* are obtained on the basis of sampling

Table 3. Meaning of the notations used in this unit.

N_h (size of the h th stratum)	total number of units in the h th stratum;
n_h	number of units selected in the sample from the h th stratum;
$W_h = \frac{N_h}{N}$ (h th stratum weight)	proportion of the population units falling in the h th stratum;
$f_h = \frac{n_h}{N_h}$	sampling fraction for the h th stratum;
Y_{hi} (y_{hi})	the value of study variable for the i th unit (or sample unit) in the h th stratum, $1 \leq i \leq N_h$;
$Y_h = \sum_{i=1}^{N_h} Y_{hi}$	h th stratum total for the estimation variable based on N_h units;
$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi}$	mean of the study variable in the h th stratum;
$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$	h th stratum sample mean for the study variable;
$\sigma_h^2 = \frac{1}{N_h} \left((\sum_{i=1}^{N_h} Y_{hi}^2) - N_h \bar{Y}_h^2 \right)$	h th stratum variance based on N_h units;
$S_h^2 = \frac{N_h}{N_h - 1} \sigma_h^2$	h th stratum mean square based on N_h units;
$s_h^2 = \frac{1}{n_h - 1} \left((\sum_{i=1}^{n_h} y_{hi}^2) - n_h \bar{y}_h^2 \right)$	sample mean square based on n_h sample units drawn from the h th stratum;

schemes followed in each individual stratum. These estimates when pooled over all the strata give an overall estimate of the respective population parameters.

Try the following exercise to get familiar with notations defined in above table.

-
- E4) Let a population of 100 units is divided into four strata of size $N_1 = 10$, $N_2 = 15$, $N_3 = 50$, $N_4 = 25$, and let the corresponding sample sizes allocation to these four strata be $n_1 = 3$, $n_2 = 4$, $n_3 = 15$, $n_4 = 8$, respectively. Also, let the value Y_{hi} of study variable for the i -th unit in the h th stratum ($1 \leq h \leq 4$) be given by $Y_{hi} = h$, $\forall i$. Calculate the value of the terms \bar{Y}_h , \bar{y}_h , σ_h^2 , S_h^2 and s_h^2 .
-

Next, we shall discuss certain relations commonly used for calculating *sample statistics* and *population estimates*. Throughout, notations will have meaning as explained in Table 3 above.

Firstly, recall that the *population mean* and the *population total* are given by relations

$$\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h, \quad \text{and} \quad Y = \sum_{h=1}^L Y_h, \text{ respectively.}$$

The following are some important relations which are used for an unbiased estimator of *population mean*, its *variance*, and the *estimated variance*.

Unbiased estimator of population mean: An estimator \bar{y}_{st} for the population mean \bar{Y} is given by

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h.$$

Variance of estimator \bar{y}_{st} : For stratified random sampling, without replacement, it is known that the sample estimator \bar{y}_{st} is unbiased and its variance is given by

$$v(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{N_h - n_h}{N_h n_h} \right) S_h^2$$

Throughout, the suffix *st* refers to *stratification*.

$$\begin{aligned}
 &= \sum_{h=1}^L W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} \\
 &= \sum_{h=1}^L W_h^2 \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h}.
 \end{aligned}$$

Estimator of variance $v(\bar{y}_{st})$: With stratified random sampling (without replacement) an unbaised estimator of the variance of \bar{y}_{st} is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{N_h - n_h}{N_h n_h}\right) s_h^2.$$

Unbiased estimation of population total: Since the population total is

$$Y = N\bar{Y} = \sum_{h=1}^L N_h \bar{Y}_h,$$

so, the (unbiased) *estimator of population total* \hat{Y} is given by

$$\hat{Y}_{st} = \sum_{h=1}^L N_h \bar{y}_h.$$

The *variance* of this estimator is given by

$$V(\hat{Y}_{st}) = \sum_{h=1}^L \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h}.$$

Unbiased estimator of proportion: As in case of simple random sampling (see Unit 12), estimation of proportion of units having an attribute is tackled by defining a variable given by

$$y_{hi} = \begin{cases} 1, & \text{if the } i\text{th unit in the } h\text{th stratum possesses the attribute} \\ 0, & \text{otherwise} \end{cases}$$

So, if N'_h denotes the number of units having the attribute in h th stratum and $P_h = \frac{N'_h}{N_h}$, then

$$N' = \sum_{h=1}^L N'_h \quad \text{and} \quad P = \frac{N'}{N} = \sum_{h=1}^L W_h P_h.$$

Also, it is clear that the population mean \bar{Y} in this case is P i.e., it equals proportion of units having the attribute. An unbiased estimator of P is given by

$$p_{st} = \sum_{h=1}^L W_h p_h, \quad \text{where } p_h = \frac{n'_h}{n_h},$$

n'_h being the number of units having the attribute in the sample in the h th stratum. Expressions for variance and estimator of variance of p_{st} are given by

$$v(p_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{N_h - n_h}{N_h - 1}\right) \left(\frac{P_h(1 - P_h)}{n_h}\right); \text{ and}$$

$$V(p_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{N_h - n_h}{N_h}\right) \left(\frac{p_h(1 - p_h)}{n_h - 1}\right)$$

In order to understand the use of some of the above given relations, let us discuss a practical problem.

Problem 1. A sample for estimating the *number of orchards of apple* is conducted in some district of Himachal Pradesh. And, *Four strata* A, B, C and D of villages are formed according to the acreage of temperate fruit trees as per records available with the revenue records. The *sizes of strata* (in acres) were 0-3, 3-6, 6-15, and 15 and above, respectively. A simple random sample of villages in each stratum was selected and the number of apple orchards was noted in selected villages. The collected data for various strata are as given below.

Stratum	Total number of villages	villages selected	No of orchards in the selected villages
A	275	15	2, 5, 1, 9, 6, 7, 0, 4, 7, 0, 5, 0, 0, 3, 0
B	146	10	21, 11, 7, 5, 6, 19, 5, 24, 30, 24
C	93	12	3, 10, 4, 11, 38, 11, 4, 46, 4, 18, 1, 19
D	62	11	30, 42, 20, 38, 29, 22, 31, 28, 66, 41, 15

Estimate the number of orchards in the district. Also estimate the variance of estimated number of orchards.

Solution. For our use in the subsequent computations, we prepare the following table of values for different quantities required in the final calculations.

Stratum-I	Stratum-II	Stratum-III	Stratum-IV
$n_1 = 15$	$n_2 = 10$	$n_3 = 12$	$n_4 = 11$
$N_1 = 275$	$N_2 = 146$	$N_3 = 93$	$N_4 = 62$
$W_1 = 0.4774$	$W_2 = 0.2535$	$W_3 = 0.1615$	$W_4 = 0.1076$
$\bar{y}_1 = 3.27$	$\bar{y}_2 = 15.2$	$\bar{y}_3 = 14.08$	$\bar{y}_4 = 32.91$
$s_1^2 = 9.6381$	$s_2^2 = 88.84$	$s_3^2 = 205.91$	$s_4^2 = 192.69$

Using values from above table, we have

$$\begin{aligned}\bar{y}_{st} &= W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3 + W_4\bar{y}_4 \\ &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3 + N_4\bar{y}_4) \\ &= \frac{1}{576}[275(3.27) + 146(15.2) + 93(14.08) + 62(32.91)] = 11.23.\end{aligned}$$

And, the estimate of variance is

$$\begin{aligned}V(\bar{y}_{st}) &= \sum_{h=1}^4 \frac{W_h^2(N_h - n_h)s_h^2}{N_h n_h} \\ &= \frac{(0.4774)^2(275 - 15)}{275 \times 15} \times 9.6381 + \frac{(0.2535)^2(146 - 10)}{146 \times 10} \times 88.84 \\ &\quad + \frac{(0.1615)^2(93 - 12)}{93 \times 12} \times 205.91 + \frac{(0.1076)^2(62 - 11)}{62 \times 11} \times 192.69 \\ &= 0.134 + 0.5321 + 0.3901 + 0.1671 = 1.2277.\end{aligned}$$

In this case, since total number of orchards in the district are to be estimated,

$$\hat{Y}_{st} = N\bar{y}_{st} = 576 \times 11.23 = 6468.48,$$

and an estimate of variance is

$$V(\hat{Y}_{st}) = (576)^2 \times 1.2277 = 407321.39.$$

————— X —————

Try the following exercise now.

- E5) Verify the values given in the last two rows of the table given in the solution of Problem 1.

As stated before, in stratified sampling, we need to address the problems of *construction of strata*, deciding the *proportion of sample* for each stratum, and calculating *sample statistics* and *population estimates*.

We start discussion with the methods of sample size allocations.

13.4 ALLOCATION OF SAMPLE SIZE

In a practical situation, the (total) sample size is normally decided by a single consideration viz., the budget available for a survey. However, the allocation of sample

size to different strata is made by a statistician. Here, it is important to remember that the precision of estimators largely depends on the *allocation plans*.

In fact, in order to increase the efficiency of estimators, it is imperative to choose a proper allocation plan. In this process, (1) the *strata sizes* i.e., the values of N_h ($1 \leq h \leq L$), (2) *variability within a stratum*, and (3) the *cost of observing a sampling unit* within various strata are three considerations which can affect the choice of allocation.

(1) *Equal allocation*, (2) *Proportional allocation*, and (3) *Optimum allocation* are the three methods of sample size allocation that are commonly used in practice. Let us discuss them one by one. In what follows, n stands for the total sample size i.e., total number of units in a stratified sample.

Equal allocation: Here, *the number of sampling units selected from each stratum is equal*. Thus, in this case,

$$n_h = \frac{n}{L}, \text{ for } h = 1, 2, \dots, L.$$

This method is preferred when strata sizes do not differ too much from each other and the information about the variability within the strata is not available. Sometimes equal allocation is also used for equal allocation of work to different strata. One of the advantage of using this method of allocation is that it is convenient for administration and field work. However, from the efficiency point of view, this is not a desirable method of sample size allocation.

Proportional allocation: This method was proposed by Bowley (1926) and has its motivation in the argument that *samples are distributed to different strata in proportion to strata sizes*. That is, larger strata should get a larger share of allocation while the smaller strata are allocated smaller number of units. Hence, here the sample allocation to the *hth-stratum* is made by

$$n_h = \frac{n N_h}{N}, \text{ where notations are used from Table 3.}$$

This method is simple to use and numerous estimates can be made with greater degree of precision by this method. However, it does not take into account an important aspect associated with stratified sampling, namely, the variability within strata.

Optimum allocation: This allocation method is given by Neyman (1934). Here, the basic idea is that, for population with larger variability, sample sizes have to be large. That is, we should take larger allocation sample sizes for strata with higher variability. Also, as we want that larger strata should have a higher allocation, so, to improve the precision of estimates (i.e., to reduce the variance), an important criterion of allocating the sample sizes should be to minimise the variance $v(\bar{y}_{st})$ of stratified sample mean estimator for a fixed total sample size, n .

The minimisation of variance $v(\bar{y}_{st})$, subject to the constraint $\sum_{h=1}^L n_h = n$, leads to the allocation

$$\begin{aligned} n_h &= n \frac{W_h S_h}{\sum_{h=1}^L W_h S_h} \\ &= n \frac{N_h S_h}{\sum_{h=1}^L N_h S_h}. \end{aligned}$$

Clearly, here we have taken into account both the strata sizes as well as strata variability.

However, in Neyman allocation as described above, it is assumed that the sampling cost per unit among different strata is same and the size of the sample is fixed.

Alternatively, if we want to minimise the cost for a specified value of the variance of stratified sample mean \bar{y}_{st} , then the simplest *cost function*, we referred to above, is

$$C = C_0 + \sum_{h=1}^L C_h n_h,$$

where C stands for the overall budget, C_0 for the (fixed) overhead cost, and C_h is the average cost of observing the study variable for each unit selected in the sample from the h th stratum. Then, an **optimum allocation** is given by that value of n_h for which C is minimum. And, using standard techniques from calculus, one can see that such a value of n_h is given by the relation

$$n_h = \frac{\frac{(C - C_0)W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^L \frac{W_h S_h}{\sqrt{C_h}}}$$

The corresponding relations for the variance, in above discussed three types of allocations methods, are given by

$$\begin{aligned} V(\bar{Y}_{st})_{eq} &= L \sum_{h=1}^L W_h^2 \left(\frac{1}{n} - \frac{1}{N_h} \right) S_h^2, \\ V(\bar{Y}_{st})_{prop} &= \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{h=1}^L W_h S_h^2, \\ V(\bar{Y}_{st})_{opt} &= \frac{1}{n} \left(\sum_{h=1}^L W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^L W_h S_h^2, \text{ respectively.} \end{aligned}$$

A comparative study between unstratified, proportional and optimum allocations shows that if we ignore *finite population correction* then

$$V(\bar{Y}_{st}) \geq V(\bar{Y}_{st})_{prop} \geq V(\bar{Y}_{st})_{opt}.$$

Thus, in terms of efficiency, optimum allocation is better than proportional allocation, which, in turn, is better than unstratified simple random sampling.

In the following problem, a practical situation is discussed to illustrate the above explained methods of sample size allocation.

Problem 2. Suppose three small towns are under study, having population

$N_1 = 50000$, $N_2 = 30000$ and $N_3 = 40000$, respectively. A stratified random sample is to be taken with a total sample size of $n = 500$. Determine the sample size to be taken from each town individually using the method of (a) proportional, and (b) optimal allocation. It is (roughly) known from a previous survey that $S_1 = 30$, $S_2 = 15$ and $S_3 = 20$. (*Notations here have same meaning as given in Table 3.*)

Solution. (a) Under proportional allocation:

$$n_1 = n \left(\frac{N_1}{N} \right) = 500 \times \frac{5}{12} = 208;$$

$$n_2 = n \left(\frac{N_2}{N} \right) = 500 \times \frac{3}{12} = 125;$$

$$n_3 = n \left(\frac{N_3}{N} \right) = 500 \times \frac{4}{12} = 167.$$

(b) Under optimal allocation:

$$n_1 = n \left(\frac{W_1 S_1}{\sum_{h=1}^3 W_h S_h} \right) = 500 \times \frac{150}{23 \times 12} = 272;$$

$$n_2 = n \left(\frac{\sum_{h=1}^3 W_h S_h}{\sum_{h=1}^3 W_h S_h} \right) = 500 \times \frac{45}{23 \times 12} = 82;$$

$$n_3 = n \left(\frac{\sum_{h=1}^3 W_h S_h}{\sum_{h=1}^3 W_h S_h} \right) = 500 \times \frac{80}{23 \times 12} = 145.$$

Now, try the following exercise for your practice.

-
- E6) In Problem 1, let the mean square errors in different strata be $S_1^2 = 8$, $S_2^2 = 2$, $S_3^2 = 15$ and $S_4^2 = 20$. Obtain the allocations using equal, proportional and optimum allocations. Also, work out the variances of the estimated mean corresponding to these allocations.
-

In this section, we discussed how to allocate the sample sizes to various strata. So, you know how many units per stratum should be selected so that the survey is cost effective and the variance of the estimate is minimum. Recall, the variability within a stratum can be minimised provided we could manage to take extra care while *constructing strata*.

Let us discuss this aspect of stratified sampling in the next section.

13.5 CONSTRUCTION OF STRATA

In the process of *construction of strata*, the basic consideration involved is that *the strata should be internally homogeneous*. For the construction of strata in which the distribution of a (single) study variable y is available, L strata could then be formed by cutting this distribution at $(L - 1)$ suitable points.

However, in practice, the distribution of y is not available always. So, in the absence of this information, the next best alternative for the formation of strata is to look at the frequency distribution of some other variable which is highly correlated with the relevant study variable y .

Here, we have to remember that the construction of strata on such an auxiliary variable will not yield exactly optimum strata, but atleast it can provide a good approximation.

Working on these lines, **Dalenius and Hodges (1957)** gave a procedure called **cumulative square root rule**. This rule is used on the frequency distribution of a highly positively correlated auxiliary variable x (also called **stratification variable**). This rule uses the argument that the distribution of y within strata can be assumed to be rectangular if the number of strata is large.

Cumulative square root rule, though proposed for the optimum allocation method, is found to yield approximately optimum strata for equal and proportional allocation methods as well. This rule can, therefore, be used for the construction of strata for all allocation methods.

The various steps involved in the construction of strata for this method are listed as under:

1. Obtain a frequency table with K classes for the **stratification variable x** .
2. In the frequency table for x , obtain square roots of the frequencies for each of the K classes.

3. Obtain the cumulative totals of the square roots of frequencies for each of the K classes. Let T denote the cumulative total for the K -th class.
4. If L strata are to be constructed, use linear interpolation method on the class intervals and cumulative square root frequency column to obtain the value of $x = x_1$, which corresponds to the value $\frac{T}{L}$ in the cumulative square root frequency column.
5. Repeat the process in above step to obtain $x = x_i$ corresponding to the value $\frac{iT}{L}$, $i = 2, 3, \dots, L - 1$, in the cumulative square root frequency column.
6. The values $(x_1, x_2, \dots, x_{L-1})$ so obtained define L strata with boundaries $(< x_1), (x_1 \text{ to } x_2), (x_2 \text{ to } x_3), \dots, (x_{L-2} \text{ to } x_{L-1}), \text{ and } (\geq x_{L-1})$.

Let us try to understand the steps involved in this rule with the help of a particular situation.

Problem 3. It is desired to estimate average annual milk yield per cow for a tharparkar herd of 127 cows at a certain government cattle farm using stratified simple random sampling. Cows in the herd are to be grouped into *three strata on the basis of first lactation length* in days. Optimum method of sample allocation is to be used for selecting the overall sample of 25 cows from the three strata. Determine approximately optimum strata boundaries using the information on first lactation length given in the table on the next page.

Lactation length	No. of cows (f)	\sqrt{f}	Cummulative \sqrt{f}
30 – 70	4	2.00	2.00
70 – 110	6	2.45	4.45
110 – 150	3	1.73	6.18
150 – 190	8	2.83	9.01
190 – 230	20	4.47	13.48
230 – 270	27	5.20	18.68
270 – 310	25	5.00	23.68
310 – 350	14	3.74	27.42
350 – 390	7	2.65	30.07
390 – 430	6	2.45	32.52
430 – 470	6	2.45	34.97
470 – 510	1	1.00	35.97

Solution. Here, we are given the frequency for the stratification variable, namely, *the first lactation length*. In the next step, we obtain the square roots of the frequencies (f) as given in the second column of above table. The third column of the gives the square root values (\sqrt{f}) and the cumulative totals of \sqrt{f} constitute the fourth column of this table.

Now, here we have $L = 3$, $K = 12$, and $T = 35.97$. For constructing three strata, we need to determine only two boundaries, x_1 and x_2 in days, using linear interpolation between the class intervals and the cumulative \sqrt{f} values. By above stated Step-4 and Step-5 of the rule, x_1 and x_2 correspond to the values

$$\frac{T}{3} = \frac{35.97}{3} = 11.99 \quad \text{and} \quad \frac{2T}{3} = \frac{2(35.97)}{3} = 23.98, \text{ respectively,}$$

in the fourth column of the table. You can see that value 9.01 in the fourth column correspond to value 190 in the first column, whereas, the value 13.48 in the fourth column corresponds to the value 230 in the first column of the table. Thus, an increase of 4.47 in cumulative \sqrt{f} value takes place over the interval 190 – 230. The first lactation length x_1 corresponding to the cumulative value of 11.99, therefore, lies in the interval 190 – 230.

Hence, by the method of *linear interpolation*,

$$11.99 = \frac{230 \times 9.01 - 190 \times 13.48}{230 - 190} + \frac{13.48 - 9.01}{230 - 190} x_1$$

Sampling

Recall that, if $f(a_1) = b_1$ and $f(a_2) = b_2$, then the value $f(x)$ of the function f at any point x on the line joining points (a_1, b_1) and (a_2, b_2) , by the method of linear interpolation, is given by the relation

$$f(x) = \frac{a_2 b_1 - a_1 b_2}{a_2 - a_1} + \frac{(b_2 - b_1)}{a_2 - a_1} x.$$

$$\Rightarrow x_1 = 216.67.$$

Similarly, it can be seen that $x_2 = 313.21$. This shows that the cows with the first lactation length in the range $[30, 216.67]$ will constitute the first stratum, whereas those having lactation length in the ranges $[216.67, 313.21]$ and $[313.21, 510]$ will form second and third strata, respectively.

Now, you try the following exercise.

- E7) It is proposed to estimate total wool yield in a certain region of Rajasthan, using stratified simple random sampling. An overall sample of 20 villages is to be selected employing Optimum method of sample allocation. The stationary sheep population data for 141 villages of this region is as in the frequency table given below.

No. of Sheep	No. of Villages (f)
0 – 100	46
100 – 200	36
200 – 300	23
300 – 400	11
400 – 500	6
500 – 600	4
600 – 700	4
700 – 800	1
800 – 900	4
900 – 1000	4
1000 – 1100	1
1100 – 1200	1

Construct three approximately optimum strata taking stationary sheep population as the stratification variable.

Next, we discuss the quantitative part of the strata formation. That is, about the method by which we decide the *number* of strata to be taken in a stratified sampling. Certainly, in any stratifying sampling, the *minimum number of strata* is two and *maximum number of strata* could be, say n , with one unit selected from each stratum.

However, to obtain an estimate of variance, one should have at least two units per stratum. Thus, the maximum number of strata may be $(n/2)$. As expected, increasing the number of strata adds towards increase in efficiency but the fact is that this *gain goes on decreasing with increase in number of strata*. Also, in situations where estimation of variance is not possible due to increase in the number of strata collapsing of strata is done.

So, as a general rule, number of strata beyond 6 to 8 is seldom profitable. It is possible to use more general and complex methods of sampling in each stratum separately. And, estimation of population parameters can be done accordingly.

Finally, in the concluding part of the unit, we want to discuss with you the concept of *post-stratification*.

13.6 POST-STRATIFICATION

Recall that, in stratified sampling, we *presuppose* that the strata sizes and the sampling frame for each stratum are available. However, there do exist situations when it is difficult to obtain these two informations.

For instance, this is a case when details about classification of farmers' population by farm size (i.e., small, medium, large) is required. Though the size of the population here can be obtained from the census records, but the *list of farmers falling in each of the three classes* may not be available. Consequently, it is not possible to determine in advance as to which stratum a farmer belongs until he is observed for the corresponding farm size.

This simply means that we can assign the units to different strata once the sample units are contacted and observed. The underlying procedure is called **post-stratification**.

The following are some useful formulations required for calculating the estimator of population mean, approximate variance, and estimator of variance when the sample units are stratified after they have been selected as a *single without replacement simple random sampling* from the entire unstratified population.

Estimator of population mean \bar{Y} : $\bar{y}_{ps} = \sum_{h=1}^L W_h \bar{y}_h$.

Approximate variance of \bar{y}_{ps} :

$$V(\bar{y}_{ps}) = \frac{N-n}{Nn} \sum_{h=1}^L W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^L (1-W_h) S_h^2.$$

Note that the first term in the last equation is the value of $V(\bar{y}_{st})_{\text{prop}}$. In fact, in a random allocation, units get distributed to different strata in proportion to strata sizes. Thus, the first term belonging to proportional allocation is as expected. And, the second term is due to deviations in proportionality and, so, is a contribution due to *post-stratification*. For large sample sizes, second term is likely to be small and post-stratification is nearly as good as proportional allocation.

Estimator of variance $V(\bar{y}_{ps})$:

$$v(\bar{y}_{ps}) = \sum_{h=1}^L \left(\frac{N_h - n_h}{N_h n_h} \right) W_h^2 S_h^2.$$

With this we end our discussion on stratified sampling. Let us summarize what we have discussed in this unit.

13.7 SUMMARY

In this unit, we have discussed the following points.

1. Basic principles of stratified sampling with the help of various type of illustrative examples. A case study of a *transport company* is analysed to facilitate the understanding of the following four questions in the context of stratified sampling.
 - a) How to form strata?
 - b) How many strata to be formed?
 - c) How to select a sampling technique for various strata?
 - d) How to allocate the sample size to different strata?
 Also, some advantages of stratified sampling method, that it has over other sampling methods have been briefly discussed.
2. With the assumption that SRS-wor is being used for sampling within a stratum, the use of the following relations has been illustrated.

$$(\text{unbiased estimator of population mean}) \bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h;$$

$$(\text{Variance of estimator } \bar{y}_{st}) v(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{N_h - n_h}{N_h n_h} \right) S_h^2;$$

$$(\text{Estimator of variance } v(\bar{y}_{st})) V(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 \left(\frac{N_h - n_h}{N_h n_h} \right) s_h^2;$$

This technique is typically useful when published journals/reports may provide clear indication of strata sizes and, due to non-availability of strata frames, it is difficult to sample the units from different strata.

The suffix **ps** here refers to *post-stratification*.

(Unbiased estimation of population total Y) $\hat{Y}_{st} = \sum_{h=1}^L N_h \bar{y}_h$;

$$(\text{variance of } \hat{Y}_{st}) V(\hat{Y}_{st}) = \sum_{h=1}^L \left(1 - \frac{n_h - 1}{N_h - 1}\right) \frac{\sigma_h^2}{n_h},$$

where the notations have their meaning as defined in Table 3.

3. The following three method of sample size allocation are explained.

a) Equal allocation $\left(n_h = \frac{n}{L}\right)$;

b) Proportional allocation $\left(n_h = \frac{n N_h}{N}\right)$;

c) optimum allocation $\left(n_h = n \frac{N_h S_h}{\sum_{h=1}^L N_h S_h}\right)$;

Also, the optimum allocation

$$n_h = \frac{\frac{(C - C_0) W_h S_h}{\sqrt{C_h}}}{\sum_{h=1}^L \frac{W_h S_h}{\sqrt{C_h}}}.$$

is discussed with reference to the cost function

$$C = C_0 + \sum_{h=1}^L C_h - h n_h.$$

4. Construction of a strata, using cummulative square root rule, is discussed. The method is illustrated with help of examples.
 5. The concept of post stratification is discussed briefly.

13.8 ANSWERS/SOLUTIONS

- E1) Do it yourself.
 E2) From the discussion held before this exercise.
 E3) Use some examples given in this unit.
 E4) Substitute values in relevant formula given in Table 3.
 E5) Use relevant formula.
 E6) Here $S_1^2 = 8, S_2^2 = 12, S_3^2 = 15$ and $S_4^2 = 20$. According to data given in Problem 1, we have

$N_1 = 275, N_2 = 146, N_3 = 93, N_4 = 62$, and total sample size $n = 48$.

(a) **Equal allocation** : $n_h = \frac{n}{4} = 12$, for all $h, 1 \leq h \leq 4$. So, the variance

$$\begin{aligned} V(\bar{y}_{st})_{eq} &= 4 \sum_{h=1}^4 W_h^2 \left(\frac{1}{48} - \frac{1}{N_h} \right) S_h^2 \\ &= 1.0034 + 0.5176 + 0.2366 + 0.0871 \\ &= 1.845. \end{aligned}$$

(b) **Propotional allocation** : Here $n_1 = 23, n_2 = 12, n_3 = 8$ and $n_4 = 5$. And so, the variance in this case

$$V(\bar{y}_{st})_{prop} = \left(\frac{1}{48} - \frac{1}{576} \right) \sum_{h=1}^4 W_h S_h^2$$

$$\begin{aligned}
 &= 0.0191[0.4774 \times 64 + 0.2535 \times 144 + 0.1615 \times 225 \\
 &\quad + 0.1076 \times 400] \\
 &= 2.797.
 \end{aligned}$$

(c) **Optimum allocation** : Here, we first compute

$$\begin{aligned}
 \sum_{h=1}^4 W_h S_h &= [3.8192 + 3.042 + 2.4225 + 2.152] \\
 &= 11.4357.
 \end{aligned}$$

Then,

$$n_1 = 48 \times \frac{3.8192}{11.4357} = 16;$$

$$n_2 = 48 \times \frac{3.042}{11.4357} = 13;$$

$$n_3 = 48 \times \frac{2.4225}{11.4357} = 10;$$

$$n_4 = 48 \times \frac{2.152}{11.4357} = 9.$$

And, the variance in this case is

$$\begin{aligned}
 V(\bar{y}_{st})_{opt} &= \frac{1}{48} \left(\sum_{h=1}^4 W_h S_h \right)^2 - \frac{1}{N} \sum_{h=1}^4 W_h S_h^2 \\
 &= 2.7245 - 0.2542 \\
 &= 2.4703.
 \end{aligned}$$

- E7) Observe that in given table we already have the frequency for the stratification variable, namely, *stationary sheep population*. For rest of the calculations follow the steps as in Problem 3.

UNIT 14 CLUSTER SAMPLING AND MULTISTAGE SAMPLING

Structure	Page No.
14.1 Introduction Objectives	44
14.2 Cluster Sampling Preliminaries Estimation of population mean Efficiency of cluster sampling	45
14.3 Multistage sampling Preliminaries Estimation of mean in two stage sampling	52
14.4 Summary	57
14.5 Answers/Solutions	57

14.1 INTRODUCTION

In all the sample selection procedures discussed so far in this block, the entire investigation was based on the assumption that a usable list of units (i.e., a *frame*) is available from which one selects a sample. Unfortunately, not always such a list of units is available especially when we are concerned with *countrywide investigations*. Even existence of such a list of units under investigation (in some cases) would not give us enough scope to base our enquiry on a simple random sample because of high budgeting costs.

For example, if the sampling units are individuals, a random sample is likely to be scattered evenly over the region under survey making it difficult to conduct survey with low cost. Alternatively, for example, if *chunks* of households or villages are selected as sampling units then these units can be *clustered* to make the survey more cost effective, particularly with respect to travel cost.

Also, as the efficiency of estimators depends on the fact whether the sampling units are clusters of individuals or the individuals themselves, the *choice of sampling unit becomes an important consideration for a proper sampling plan*. And, the problem of frames can be handled more effectively by forming *clusters of basic sampling units*.

The procedure of selecting **clusters** and then observing all the elements in the selected clusters is known as **cluster sampling**. A natural extension of idea of cluster sampling is **sub-sampling** in which the clusters selected at a later stage are further sub-sampled. The procedure of sub-sampling can be extended to **multi-stage sampling**.

In Sec.14.2, we shall talk about certain preliminary aspects of *cluster sampling*, discuss relations used in the *estimation of population mean*, and describe briefly the *efficiency of cluster sampling*. In Sec.14.3, after having discussed certain preliminary aspects of *multi-stage sampling*, we shall discuss another set of relations used in the estimation of population mean in the context of *two-stage sampling*.

Objectives

After reading this unit, you should be able to

- discuss a situation for using cluster/multistage sampling;
- estimate the population mean in case of equal and unequal size of clusters;
- estimate the relative efficiency of cluster sampling;

- differentiate between cluster sampling and two-stage sampling;
- estimate the population mean in case of two-stage sampling.

14.2 CLUSTER SAMPLING

As said in the introduction, when the sampling unit is a cluster, the procedure of sampling is called cluster sampling. So, **cluster sampling** consists of forming suitable clusters of contiguous population units and surveying all the units in a sample of clusters selected according to some appropriate sample selection method.

For instance, consider a big village as a cluster of farmers. Then, for selecting farmers from the area, certain smaller villages may be selected and information from farmers of these villages is obtained. Here, it is important to mention that a *list of farmers* in a region may not be available but a *list of villages* is always available. This example is typical in *area sampling*.

Other than the *area sampling*, there are situations when cluster sampling is of great help. For example, if one wishes to interview passengers departing from an airport, then a *cluster might be the plane load*. On the other hand, if one is searching through files of land holdings for tax information, then *pages in a ledger* would be the *clusters*.

There are situations when conducting a survey with clusters of sampling units, instead of taking a simple random sample from a population, is cost effective.

For instance, if a sample is selected from the population of all *sixth-grade* students in a particular state, then each school in the state is taken as a cluster of the basic sampling units and we choose a simple random sample of a few schools and interview all the *sixth-graders* in those schools according to pre-set survey objectives. However, a simple random sample of 400 students usually, as we may agree to, better represents the entire population – and therefore provides better information about the population – than a group of 100 students studied in each of the four specified schools.

Thus, in general, the choice for a sampling procedure to be adopted should be guided by cost considerations and by the degree of precision desired in estimating the population parameters.

Also, it is important to realise that the cluster sampling in above situation has actually helped in avoiding the necessity of constructing a *frame* for the entire population, which is certainly an exhausting and expensive job in itself. In addition to that, cluster sampling is remarkably expedient because the units in a cluster are adjacent and therefore easy to approach.

Now, let us talk about some more aspects of cluster sampling in detail. We start by talking about some introductory aspects of cluster sampling.

14.2.1 Preliminaries

Let us consider a case of cluster sampling in which a number of people in a city are to be interviewed. For selecting a sample, the telephone directories are used and it is decided to interview people through telephone. Now, since all the residents can be numbered, a random sampling technique could have been used to choose sample houses.

Also, we could form strata of houses for *high*, *middle*, and *low* income groups. Now, if we choose houses throughout the city in random manner, then cost of visiting widely scattered dwellings will certainly be prohibitive.

An alternative way of sample selection is to *group blocks* or *areas* into clusters of approximately equal population. Then, a number of these clusters can be chosen at

random. Within each cluster, all households may be interviewed. On comparing this (cluster) sampling procedure with that of making random choice of households throughout the city, it is clear that the cost per element (a household) is certainly going to be lower because of lower listing cost (as it is necessary only to list the houses on the blocks selected) and lower location cost. Also, it is going to be easy for an interviewer to talk to several people on one block rather than to several people scattered throughout the city.

Note that a necessary condition for the validity of above procedure is that every unit of the population under study must correspond to one and only one unit of the cluster so that the total number of sampling units in the list (frame) will cover all the units of the population under study with no omission or duplication. When this condition is not satisfied, estimators become **biased**.

In the following table, we fix some notations for our convenience and future use in this unit. We shall use these notations frequently while calculating estimators of population parameters.

Table 1. Notations used in this section and their meanings.

N n $M_i, 1 \leq i \leq N$ $M_0 = \sum_{i=1}^N M_i$ $\bar{M} = \frac{M_0}{N}$ Y_{ij} $Y_i = \sum_{j=1}^{M_i} Y_{ij}$ $Y = \sum_{i=1}^N Y_i = \sum_{i=1}^N \sum_{j=1}^{M_i} Y_{ij}$ $\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} Y_{ij} = \frac{Y_i}{M_i}$ $y_k = \sum_{k=1}^{M_k} Y_{kj} (= Y_k)$ $\bar{Y}_c = \frac{Y}{N} = \frac{1}{N} \sum_{i=1}^N Y_i$ $\bar{Y} = \frac{\sum_{i=1}^N M_i \bar{Y}_i}{M_0} = \frac{Y}{M_0}$	number of clusters in a population. number of clusters in the sample. number of units in the i -th cluster of a population. total number of units in a population. average number of units per cluster in a population. value of the character under study for the j -th unit in the i -th cluster, $j = 1, 2, \dots, M_i; i = 1, 2, \dots, N$. <i>i</i> -th cluster total. total of Y -values for all the M_0 units in a population. the mean per unit of the i -th cluster. k -th sample cluster total ($1 \leq k \leq n$). population mean per cluster. population mean per unit.
---	---

Let us workout a problem to get familiar to the notations defined in Table 1.

Problem 1. Suppose from a total of 20 bearing trees of guava in a village, 5 clusters of size 4 trees each were selected and (hypothetical) yield (in kgs) is as given in the following table.

cluster	1st tree	2nd tree	3rd tree	4th tree
1	5	4	1	15
2	11	1	4	7
3	36	10	19	11
4	7	15	12	10
5	2	22	8	6

Calculate the quantities Y_i , \bar{Y}_i ($1 \leq i \leq 5$), \bar{Y}_c and \bar{Y} .

Solution. Here, $n = 5$, $M_i = 4$, for all i , and $N = 20$. Then, using values of Y_{ij} from above table, we get

$$Y_1 = \sum_{j=1}^{M_i} Y_{1j} = \sum_{j=1}^4 Y_{1j} = 5 + 4 + 1 + 15 = 25, \text{ and}$$

$$Y_2 = \sum_{j=1}^4 Y_{2j} = 11 + 1 + 4 + 7 = 23.$$

Similarly, you can find that $Y_3 = 76$, $Y_4 = 44$ and $Y_5 = 38$. Thus,

$$\bar{Y}_1 = \frac{Y_1}{M_1} = \frac{25}{4}, \quad \bar{Y}_2 = \frac{Y_2}{M_2} = \frac{23}{4}, \quad \bar{Y}_3 = \frac{76}{4}, \quad \bar{Y}_4 = \frac{44}{4} \text{ and } \bar{Y}_5 = \frac{38}{4}.$$

Finally,

$$\bar{Y}_c = \frac{1}{N} \sum_{i=1}^N Y_i = \frac{1}{5} (25 + 23 + 76 + 44 + 38) = 41.2, \text{ and}$$

$$\bar{Y} = \frac{Y}{M_0} = \frac{\sum_{i=1}^5 Y_i}{M_0} = \frac{25 + 23 + 76 + 44 + 38}{20} = 10.3.$$

————— × —————

Now you try to solve the following exercise.

- E1) From your daily life experiences, give five examples where cluster sampling can be used.

As you may have observed, the clusters are usually formed by grouping neighbouring units or units which can be conveniently surveyed together. The construction of clusters, however, differs from the optimal construction of strata that you read about in Unit 13.

For example, household is a cluster of individuals, village is a cluster of farmers and a class, which is a group of students, is a cluster. Similarly, an orchard can be considered as cluster of trees, etc.

Generally, while stratification may reduce sampling error, clustering tends to decrease costs and increase sampling error for the same size of sample. This is mainly because people who live close together are more likely to be similar than those living in different localities.

As you read in previous unit, while stratifying a population, strata are as homogeneous as possible within themselves and differ as much as possible from each another with respect to the study variable. And, units within a stratum need not be geographically contiguous. On the other hand, if an estimate based on cluster sampling has to be more efficient than the estimates made with simple random sampling procedure, clusters should be internally as *heterogeneous* as possible.

For a given total number of units in the sample, the cluster sampling is usually less efficient than sampling of individual units as the latter is likely to provide a better cross section of the population units than the former. This is essentially due to tendency of units in a cluster to be similar. Another fact that we would like to share with you is that the efficiency of cluster sampling is likely to decrease with increase in cluster size.

All said and done, cluster sampling is operationally convenient and economical than

Sampling

Some of the major Government agencies, Universities research studies, and marketing research use a combination of clustering and stratification techniques to control cost and error and to provide adequate size groups for intensive studies.

sampling of individual units. In many practical situations, the loss in efficiency in terms of sampling variance is likely to be balanced by the reduction in cost particularly the travel cost between units.

Hence, because of its operational convenience and possible reduction in cost, the survey tasks in many situations are facilitated by using nonoverlapping and collectively exhaustive cluster of units. Now, in the next part of the section, we shall discuss some relations used in the estimation of population parameters.

14.2.2 Estimation of Population Mean

Simple random sampling, systematic sampling, and stratified sampling are various types of sampling procedures that can be applied in the cluster sampling by treating the clusters as sampling units. In this unit, however, we shall restrict to situations where clusters are selected using *without replacement simple random sampling* procedure.

The theory of cluster sampling in its own right is rather complex, where the complexity depends on whether one takes *equal* or *unequal-sized* clusters. In general, a formulae for calculating the *standard error* of cluster estimates has two terms, where the first relates to the variability between cluster means (or proportions) and the second to the variability within cluster.

In this unit, we start with the case of unequal clusters and then deduce from this the results about clusters of equal sizes as a special case.

Case-I: Unequal clusters. Usually, in practice, clusters are of unequal sizes. For instance, households as a group of persons and villages as a group of households can be taken as clusters for the purpose of sampling.

We assume (i) the population consists of N clusters, where the i th cluster has M_i elements, $i = 1, 2, \dots, N$, and (ii) n clusters are selected from N clusters by *without replacement simple random sampling* procedure.

Then, an unbiased estimator of population mean \bar{Y} , with M_0 known, is given by the relations

$$\begin{aligned}\hat{Y}_c &= \frac{N}{nM_0} \sum_{k=1}^n M_k \bar{Y}_k \\ &= \frac{1}{\bar{M}n} \sum_{k=1}^n Y_k, \text{ where } \bar{M} = \frac{M_0}{N}.\end{aligned}$$

And, the variance of the estimator \hat{Y}_c is given by the relation

$$V(\hat{Y}_c) = \left(\frac{N-n}{N n \bar{M}^2} \right) \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y}_c)^2.$$

Also, an unbaised estimator of variance is given by relations

$$\begin{aligned}\hat{V}(\hat{Y}_c) &= \left(\frac{N-n}{N n \bar{M}^2} \right) \frac{1}{n-1} \sum_{k=1}^n (Y_k - \bar{M} \hat{Y}_c)^2 \\ &= \left(\frac{N-n}{N n \bar{M}^2} \right) \frac{1}{n-1} \left[\left(\sum_{k=1}^n Y_k^2 \right) - n(\bar{M} \hat{Y}_c)^2 \right]\end{aligned}$$

We try to understand a use of these relations with the help of the following problem.

Problem 2. For studying the cultivation practices and yield of apple, a pilot sample survey is conducted in a district of Kashmir. The yield (in kgs) of 3 clusters of trees, selected by *without replacement simple random sampling*, from 15 are as given in the following table.

cluster	size	yield
1	12	5.53, 26.11, 11.08, 12.66, 0.87, 6.40, 54.31, 37.94, 7.13, 3.53, 14.23, 1.24
2	10	4.84, 10.93, 0.65, 32.52, 3.56, 11.68 35.97, 47.07, 17.69, 40.7
3	6	15.79, 11.18, 27.54, 28.11, 21.70, 1.25

With $\bar{M} = 10$, estimate the average yield per tree as well as the production of apple in the village and their standard errors.

Solution. Here, $N = 15$, $n = 3$ and $\bar{M} = 10$. Then, $M_0 = N\bar{M} = 150$. So, $M_1 = 12$, $M_2 = 10$ and $M_3 = 6$. Then, using values from table, we get

$$\begin{aligned}\hat{Y}_c &= \frac{1}{30} [Y_1 + Y_2 + Y_3] = \frac{1}{30} \left[\sum_{j=1}^{M_1} Y_{1j} + \sum_{j=1}^{M_2} Y_{2j} + \sum_{j=1}^{M_3} Y_{3j} \right] \\ &= \frac{1}{30} [181.03 + 205.61 + 105.57] = \frac{492.21}{30} = 16.407.\end{aligned}$$

Thus, the average yield of apple per tree is 16.407 kgs. Also, we have

$$Y_1 = \sum_{j=1}^{M_1} Y_{1j} = 181.03, \quad Y_2 = \sum_{j=1}^{M_2} Y_{2j} = 205.61, \quad \text{and} \quad Y_3 = \sum_{j=1}^{M_3} Y_{3j} = 105.57.$$

Using these values of Y_i ($1 \leq i \leq 3$), the estimated variance $\hat{V}(\hat{Y}_c)$ is given by

$$\begin{aligned}\hat{V}(\hat{Y}_c) &= \left(\frac{15 - 3}{15 \times 3 \times 100} \right) \frac{1}{3-1} \left[\left(\sum_{k=1}^n Y_k^2 \right) - 3(10 \times 19.709) \right] \\ &= \frac{1}{750} [(32771.86 + 42275.47 + 11145.03) - 26918.97] = 79.03\end{aligned}$$

Thus the standard error of \hat{Y}_c is $\sqrt{79.03} = 8.89$.

————— X —————

Case-II: Equal clusters. Let $M_i = M$, for all i . That is, unequal clusters reduce to clusters of equal sizes. Hence, in case of equal clusters, the unbiased estimator of population mean is given by

$$\hat{Y}_c = \frac{1}{n} \sum_{k=1}^n \bar{Y}_k, \text{ as } M_0 = NM, \text{ in this case.}$$

Similarly, the variance of the estimator \hat{Y}_c is now given by

$$\begin{aligned}V(\hat{Y}_c) &= \left(\frac{1}{n} - \frac{1}{N} \right) S_b^2 \\ &\cong \left(\frac{1}{n} - \frac{1}{N} \right) S^2 \{1 + (M - 1)\rho\}, \text{ for large } N,\end{aligned}$$

where ρ is *intra-cluster correlation coefficient* between elements within clusters, S^2 is population mean square and (with $\bar{M} = M$)

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2.$$

ρ is a measure of *internal homogeneity* of the clusters.

Generally, ρ is positive since clusters are usually found by putting together geographical contiguous farms, stores, establishments, families, etc. Thus, for the same number of units in a sample, cluster sampling gives a higher variance than sampling elements directly.

But the real point here is that it is far cheaper to collect information on a *per-unit basis* if sampling is done in clusters. If ρ is negative, both cost and the variance suggest a use of clusters. Furthermore, an unbiased estimator of $V(\hat{Y}_c)$ in this situation is given by

$$\hat{V}(\hat{Y}_c) = \left(\frac{1}{n} - \frac{1}{N} \right) s_b^2, \text{ where } s_b^2 = \frac{1}{n-1} \sum_{k=1}^n (\bar{Y}_k - \hat{Y}_c)^2$$

Once again, we shall try to understand these relations with help of a practical situation.

Problem 3. A pilot sample survey was conducted to study the management practices and yields of apple in a village of Himachal Pradesh (India). Of the total 300 bearing trees, 10 clusters of size 3 each were selected and their yield records (in kgs) are as given in the following table.

cluster	1st tree	2nd tree	3rd tree
1	6.52	5.73	15.24
2	12.08	1.65	9.28
3	24.15	30.75	17.26
4	16.24	8.20	6.58
5	54.92	34.62	12.16
6	36.24	46.28	28.54
7	42.48	36.34	26.42
8	18.24	16.80	91.46
9	54.27	40.28	25.55
10	1.94	5.16	22.70

Estimate the average yield per tree along with its standard error.

Solution. Here, $M_0 = 300$, $n = 10$ and $M_i = 3$, for $i = 1, 2, \dots, 10$. So, proceeding as in Problem 2, you can see that the average yield per tree is $\hat{Y}_c = 24.94$. Also, the estimated variance $\hat{V}(\hat{Y}_c)$, with $N = \frac{M_0}{M} = 100$, is given by

$$\begin{aligned}
 \hat{V}(\hat{Y}_c) &= \left(\frac{1}{10} - \frac{1}{100} \right) s_b^2 = \frac{9}{100} \times \frac{1}{9} \times \left(\sum_{k=1}^{10} (\bar{Y}_k - \hat{Y}_c)^2 \right) \\
 &= \frac{1}{100} \left[\left(\frac{Y_1}{M_1} - \hat{Y}_c \right)^2 + \left(\frac{Y_2}{M_2} - \hat{Y}_c \right)^2 + \dots + \left(\frac{Y_{10}}{M_{10}} - \hat{Y}_c \right)^2 \right] \\
 &= \frac{1}{9 \times 100} \left[(Y_1 - 3\hat{Y}_c)^2 + (Y_2 - 3\hat{Y}_c)^2 + \dots + (Y_{10} - 3\hat{Y}_c)^2 \right] \\
 &= \frac{1}{900} [(27.49 - 74.81)^2 + (23.01 - 74.81)^2 + \dots + (29.8 - 74.81)^2] \\
 &= 18.399.
 \end{aligned}$$

Thus, the standard error is $\sqrt{18.399} = 4.299$.

————— X —————

Now, you try the following exercise.

- E2) Change some of the figures in the last three columns of the table given in Problem 3 above, and then calculate the average yield per tree along with its standard error.

Above we obtained expressions for an unbiased estimator of population mean alongwith expressions for its variance and estimator of variance. In the next part of the section, we shall describe the relative efficiency aspect of cluster sampling. For this purpose, we shall assume situations where all clusters are of equal size.

14.2.3 Efficiency of Cluster Sampling

Here, right in the beginning, we want to remark that the estimator \hat{Y}_c for equal sized clusters is based on a sample of nM units in the form of n clusters each consisting of M units. Thus, if the same number of units are selected from a population of NM units by *without replacement simple random sampling* procedure, then the sample mean estimator \hat{Y} and its variance $V(\hat{Y})$ are given by the relations (see Unit 12)

$$\hat{Y} = \frac{1}{nM} \sum_{k=1}^{nM} Y_k, \text{ and}$$

$$\begin{aligned} V(\hat{\bar{Y}}) &= \left(\frac{1}{nM} - \frac{1}{NM} \right) S^2 \\ &= \left(\frac{N-n}{N n M} \right) \frac{1}{NM-1} \left(\left(\sum_{i=1}^{NM} Y_i^2 \right) - NM \bar{Y}^2 \right), \text{ respectively.} \end{aligned}$$

And, in relation to the sample mean estimator $\hat{\bar{Y}}$, the *relative efficiency* (RE, in short) of the estimator $\hat{\bar{Y}}_c$ for equal sized clusters is given by $RE = \frac{V(\hat{\bar{Y}})}{V(\hat{\bar{Y}}_c)}$, where $V(\hat{\bar{Y}}_c)$ denotes the variance for equal sized cluster.

Observe that the relative efficiency defined above involves value of study variable for all population units. However, in practice, the investigator has only the sample observations of n clusters of M units each. For this, he needs the estimates of two variances involved in the formulae of relative efficiency (RE).

An unbiased estimator of $V(\hat{\bar{Y}})$ from a cluster sample is given by

$$\hat{V}(\hat{\bar{Y}}) = \frac{N-n}{(NM-1)n} \left[\left(\frac{1}{nM} \sum_{k=1}^n \sum_{j=1}^M y_{kj}^2 \right) + \hat{V}(\hat{\bar{Y}}_c) - \hat{\bar{Y}}_c^2 \right],$$

while an unbiased estimator of $V(\hat{\bar{Y}}_c)$ for equal size clusters is same as given in the previous part of the section. Then, *estimator of relative efficiency* \hat{RE} of estimator $\hat{\bar{Y}}_c$ (for equal size clusters) with respect to the usual estimator $\hat{\bar{Y}}$ from a cluster sample is given by $\hat{RE} = \frac{\hat{V}(\hat{\bar{Y}})}{\hat{V}(\hat{\bar{Y}}_c)}$.

Let us now discuss a practical situation to see how above stated relation are used in practice.

Problem 4. A company has 25 centres located at different places in a State. Each centre has been provided with 4 telephones. In order to estimate the average number of calls per telephone made on a typical day for this company, a sample of 5 centres, using without replacement simple random sampling, were selected. The data regarding the number of calls made on a typical working day from each telephone of the sample centres are as summarized in Table 2.

Table 2 : Number of calls made from selected centres.

Centre	M	Calls made				\bar{Y}_i	\bar{Y}_i
1	4	26	34	27	25	112	28
2	4	44	33	28	31	136	34
3	4	18	33	25	28	104	26
4	4	37	21	22	40	120	30
5	4	23	34	42	29	128	32

Estimate the average number of daily calls per telephone made from all the 25 centres. Also, estimate the relative efficiency of the estimator used with respect to the usual sample mean estimator, from the sample selected above.

Solution. Here, $N=25$, $n=5$ and $M=4$. The sample cluster means are given in the last column of Table 2. The estimate of average number of daily calls can be computed using estimates $\hat{\bar{Y}}_c$ for equal clusters. Using values from the last column of Table 2, we have

$$\hat{\bar{Y}}_c = \frac{1}{n} \sum_{i=1}^n \bar{Y}_i = \frac{1}{5} (28 + 34 + 26 + 30 + 30) = 30.$$

Also, since the variance estimator $\hat{V}(\hat{\bar{Y}}_c)$ for equal clusters is given by the relation

$$\hat{V}(\hat{\bar{Y}}_c) = \frac{N-n}{N n(n-1)} \left(\sum_{i=1}^n \bar{Y}_i^2 - n \hat{\bar{Y}}_c^2 \right),$$

so, on making substitution, we get

$$\hat{V}(\hat{\bar{Y}}_c) = \frac{25 - 5}{(25)(5)(4)} [(28)^2 + (34)^2 + \dots + (32)^2 - 5(30)^2] = 1.6.$$

Now, we know that the variance estimator of the sample mean estimator $\hat{\bar{Y}}$ is given by the relation

$$\hat{V}(\hat{\bar{Y}}) = \frac{N - n}{(NM - 1)n} \left[\frac{1}{nM} \sum_{i=1}^n \sum_{j=1}^M Y_{ij}^2 + \hat{V}(\hat{\bar{Y}}_c) - \hat{\bar{Y}}_c^2 \right].$$

To make calculations easy, let us first compute the term involving sum of squares of all the individual observations. Once again, using values from above table, we get

$$\sum_{i=1}^n \sum_{j=1}^M Y_{ij}^2 = (26)^2 + (34)^2 + \dots + (29)^2 = 18962.$$

Thus, by above stated relations,

$$\hat{V}(\hat{\bar{Y}}) = \frac{25 - 5}{[(25)(4) - 1]5} \left[\frac{189621}{(5)(4)} + 1.6 - (30)^2 \right] = 2.0081.$$

Finally, the estimate of percent relative efficiency will be

$$\hat{RE} = \frac{\hat{V}(\hat{\bar{Y}})}{\hat{V}(\hat{\bar{Y}}_c)} \times 100 = \frac{2.0081}{1.6} (100) = 125.5.$$

So, we can infer that here cluster sampling of centres is more efficient than the usual sample mean estimator when individual telephones would have been selected.

Now you try the following exercise.

- E3) Suppose from a total of 120 bearing trees of guava in a village, 5 clusters of 4 trees each are selected and (hypothetical) yield (in kg) recorded is as given in the following table.

Cluster	1 st tree	2 nd tree	3 rd tree	4 th tree
1	5	4	1	15
2	11	1	4	7
3	26	10	19	11
4	7	15	12	10
5	2	22	8	6

Estimate average yield (in kg) per tree of guava along with its standard error.
Also, estimate the relative efficiency of the estimator.

In this section, we discussed sampling procedures in which all the elements of the selected clusters were enumerated. This scheme, as you may have observed, is convenient and economical but the method restricts the spread of the sample over a population which generally reduces the efficiency of the estimator.

We now turn to the situation in which we first select the clusters and then randomly choose a specified number of units from clusters selected before. This procedure is known as **two-stage sampling or sub-sampling**.

14.3 MULTISTAGE SAMPLING

From previous section, recall the example of cluster sampling in which we grouped blocks of a city into clusters of approximately equal population. Suppose, instead of interviewing all households in sample clusters, we make a *random choice of households within each sample clusters*. That is, a sample is now selected in two stages – first

select a sample of clusters, called *first-stage* or **primary sampling units**, and then select a sample of elements within sample clusters.

We have the following three advantages of this sampling procedure.

- (1) Lists have to be prepared for the selected *primary sampling units* and subsequent stage units only;
- (2) It is easy to check the correctness of the list; and
- (3) A sample gets concentrated in the selected *primary sampling units* and this reduces costs of travel, etc.

In this course, we shall discuss two-stage sampling. In Table 3 below, we define notations that we shall need while discussing with you various relations used in the two-stage estimation procedure.

Table 3. Notations used in this section and their meanings.

N	number of <i>primary stage units (psu's)</i> in a population
n	number of <i>psu's selected in the sample</i>
M	number of <i>second stage units (ssu's)</i> in each <i>psu</i>
m	number of <i>ssu's selected from M ssu's</i>
$M_0 = N M$	total number of <i>ssu's in a population</i>
Y_{ij}	value of the study variable y for j th <i>ssu</i> of the i th <i>psu</i> , $j = 1, 2, \dots, M$; $i = 1, 2, \dots, N$
y_{ij}	value of the study variable for j th selected <i>ssu</i> of the i th selected <i>psu</i> , $j = 1, 2, \dots, m$; $i = 1, 2, \dots, n$
$\bar{Y}_i = \sum_{j=1}^M Y_{ij}$	total of <i>Y</i> -values for the i th <i>psu</i>
$\bar{Y} = \sum_{i=1}^N \bar{Y}_i$	population total of <i>Y</i> -values
$\bar{Y}_i = \frac{\bar{Y}_i}{M}$	population mean for i th <i>psu</i>
$\bar{Y} = \frac{1}{N} \sum_{i=1}^N \bar{Y}_i$	population mean for the study variable
$y_i = \sum_{j=1}^m y_{ij}$	sample total for the i th <i>psu</i>
$y = \sum_{i=1}^n y_i$	total of <i>y</i> -values for the whole sample
$\bar{y}_i = \frac{y_i}{m}$	sample mean for the i th <i>psu</i>

To familiarize yourself with these notations, try the following exercise.

-
- E4) There are 50 fields in a village sown with wheat and each is divided into 8 plots of equal size. Out of 50 fields, 5 are selected by *without replacement simple random sampling* method. Again, from each selected field, 2 plots are chosen by *without replacement simple random sampling* method. The yield in kg/plot recorded is as given in the following table.

Selected field	Plot-1	Plot-2
1	4.16	4.76
2	5.40	3.52
3	4.12	3.73
4	4.38	5.67
5	5.31	2.59

Estimate the quantities \bar{Y} and \bar{y}_i .

14.3.1 Preliminaries

As said above, in a *two-stage* sampling design, sample clusters form the units of sampling at the first stage, called *primary stage units* (*psu's*, in short). Then, the elements within clusters are called *second stage units* (*ssu's*, in short). It is now clear that this procedure can be generalised to three or more stages and that is why it is called **multi-stage sampling**.

For example, in a survey for estimating yield of a crop, a block in a district may be taken for *primary stage units*, villages within blocks as *second stage units*, the crop fields within village as *third-stage units*, and a plot of specified shape and size within field as the *ultimate unit* of sampling.

Multistage sampling has been found to be very useful in practice and is commonly used in large-scale surveys. This sampling procedure is a compromise between cluster sampling and direct sampling of units. Furthermore, this design is more flexible as it permits the use of different sample selection procedures at different stages.

It is important to mention here that multi-stage sampling is only choice in a number of practical situations, especially when a satisfactory sampling frame of ultimate-stage units is not readily available and cost of obtaining this information is large and time consuming.

In a *multi-stage sampling* procedure, the basic idea used in the estimation of population parameters is that of building up estimates from the bottom (last stage units) to the top. It is desirable that you keep this principle in mind while reading through the next part of the section.

14.3.2 Estimation of Population Mean

As we said before, only *two stage sampling* procedure will be discussed in this unit. Also, we shall assume that the first stage units are of equal size and that units at the first and second stage are selected by *without replacement simple random sampling* procedure.

Now to select a sample, we use a frame listing all the N *psu's* in the population. A *without replacement simple random sample* of n *psu's* can then be drawn using procedure as described in Unit 12. So, a frame listing all the M second stage units in i -th selected *psu* ($i = 1, 2, \dots, n$) is obtained. Finally, a without replacement simple random sample of m units is drawn from the i -th selected *psu*, $1 \leq i \leq n$, containing M second stage units.

Then, the sample mean in this situation is given by

$$\hat{Y}_{2s} = \frac{1}{n m} \sum_{i=1}^n \sum_{j=1}^m y_{ij} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i.$$

This relation is based on all the nm units in the sample and is an unbiased estimator of population mean \bar{Y} .

For example, suppose in some locality of a city there are N *mohalla's* and we select one *mohalla* at random. Let the selected *mohalla* contains M households out of which m are selected at random. Now, we collect information on y (e.g., let it be weekly expenditure on fruits) from every household in the sample. Then, the sample mean \bar{y}_1 estimates the average (expenditure per household) in the *mohalla* and $M\bar{y}_1$ estimates the total (expenditure on fruits) for the whole *mohalla*. But, since this *mohalla* was selected at random from a total of N in the locality, the estimate of the total in the locality is $NM\bar{y}_1$.

Now suppose that not 1 but n *mohalla's* from N were selected without replacement with equal probabilities and each selected *mohalla* contain M household. Now, at the second stage of the sampling, we take a random sample of m households from each *mohalla*.

Throughout, the suffix 2s refers to **two-stage sampling** procedure.

Then, as before, $M\bar{y}_i$ ($i = 1, 2, \dots, n$) gives the estimate of the total (expenditure on fruits) for the i th *mohalla*. Therefore, $M \sum_{i=1}^n \bar{y}_i$ will give estimate for the total in the sample *mohalla's*. Hence,

$$\hat{\bar{Y}} = \frac{1}{nM} \left(M \sum_{i=1}^n \bar{y}_i \right) = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

provides the estimate for the total expenditure on fruits for the whole locality.

Again, the variance of the estimator $\hat{\bar{Y}}_{2s}$ is given by the relation

$$V(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n} - \frac{1}{N} \right) S_b^2 + \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M} \right) S_w^2, \text{ where}$$

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{\bar{Y}})^2 \text{ and } S_w^2 = \frac{1}{N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2$$

Furthermore, an unbiased estimator of $V(\hat{\bar{Y}}_{2s})$ is given by the relations

$$\hat{V}(\hat{\bar{Y}}_{2s}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_b^2 + \frac{1}{N} \left(\frac{1}{m} - \frac{1}{M} \right) s_w^2$$

$$= \left(\frac{N-n}{Nn} \right) s_b^2 + \frac{1}{N} \left(\frac{M-m}{NM} \right) s_w^2, \text{ where}$$

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{\bar{Y}}_{2s})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n \bar{y}_i^2 - n \hat{\bar{Y}}_{2s}^2 \right] \text{ and}$$

$$s_w^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2$$

$$= \frac{1}{n(m-1)} \left[\sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 - m \sum_{i=1}^n \bar{y}_i^2 \right]$$

Observe that in the expression for the variance of the sample mean $\hat{\bar{Y}}_{2s}$ there are two components wherein the first represents the contribution arising from sampling of first-stage units and the second is arising from subsampling within the selected first-stage units.

It is important to note the following two special cases of two stage sampling procedure.

- (i) $n = N$, corresponds to stratified sampling with N first stage units as strata and m units drawn from each stratum; and
- (ii) $m = M$, corresponds to cluster sampling.

Let us now consider a practical situation working of which will help us understand the use of above stated relations.

Problem 5. Assume that in Problem 4, from each selected centres, 2 telephones were chosen by *without replacement simple random sampling* method. The following table gives the data related to the number of calls made on a typical working day from chosen telephone of the selected centres.

Center	M	m	Calls made	y_i	\bar{y}_i
1	4	2	26	34	60
2	4	2	44	28	72
3	4	2	33	25	58
4	4	2	37	21	58
5	4	2	23	29	52

Estimate the average number of daily calls per telephone made from all the 25 centres alongwith its estimate of variance.

Solution. Here, $N = 25$, $n = 5$, $M = 4$, and $m = 2$. Also, observe that the sample means for selected cluster (psu) is as given in the last column of the table. Now, the estimate of average number of daily calls can be computed using estimator $\hat{\bar{Y}}_{2s}$. Thus,

using values from the last column of above table, we get

$$\hat{Y}_{2s} = \frac{1}{5} \sum_{i=1}^5 \bar{y}_i = \frac{1}{5} [30 + 36 + 29 + 29 + 26] = 30.$$

Also, we know that the estimated variance of \hat{Y}_{2s} is given by the relation

$$\hat{V}(\hat{Y}_{2s}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_b^2 + \frac{1}{N} \left(\frac{1}{m} - \frac{1}{M} \right) s_w^2.$$

Here, using values from table and the value of \hat{Y}_{2s} , we get

$$\begin{aligned} s_b^2 &= \frac{1}{n-1} \left[\sum_{i=1}^n \bar{y}_i^2 - n \hat{Y}_{2s}^2 \right] \\ &= \frac{1}{5-1} [(30)^2 + (36)^2 + \dots + (26)^2 - 5(30)^2] \\ &= \frac{1}{4}[4554 - 4500] = 13.5. \end{aligned}$$

Next, we calculate s_w^2 . For, we first compute the term involving sum of squares of all the individual observations. Thus, using values from above table, we get

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 &= (26)^2 + (34)^2 + \dots + (29)^2 = 9446, \text{ and} \\ \sum_{i=1}^n \bar{y}_i^2 &= (30)^2 + (36)^2 + \dots + (26)^2 = 4554. \end{aligned}$$

Thus,

$$\begin{aligned} s_w^2 &= \frac{1}{n(m-1)} \left[\sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 - m \sum_{i=1}^n \bar{y}_i^2 \right] \\ &= \frac{1}{5(2-1)} [9446 - 2 \times 4554] \\ &= \frac{1}{5}[338] = 67.6. \end{aligned}$$

Hence, the estimate of variance is

$$\begin{aligned} \hat{V}(\hat{Y}_{2s}) &= \left(\frac{N-n}{Nn} \right) S_b^2 + \frac{1}{N} \left(\frac{M-m}{Mm} \right) S_w^2 \\ &= \left(\frac{25-5}{25 \times 5} \right) (13.5) + \frac{1}{25} \left(\frac{4-2}{4 \times 2} \right) (67.6) \\ &= \left(\frac{20}{125} \right) (13.5) + \left(\frac{2}{200} \right) (67.6) = 2.16 + 0.676 = 2.836. \end{aligned}$$

It is generally observed that in the variance expressions given above, the contributions due to first stage sampling is much larger than the contributions due to second stage.

So, while estimating the variance, it may be approximated by the expression $\frac{s_b^2}{n}$.

A distinct advantage of multistage sampling is that the sampling variance may be broken up into as many components as there are stages. The expressions for unequal psu's are also available. However, we have already decided to consider only the case of equal psu's. The concept of multistage sampling is so common that it is difficult to visualise a real life survey situation where it has not been used.

Solve the following exercise.

-
- E5) Assume that in E3, from each selected cluster, 2 trees were chosen by *without replacement simple random sampling* method. Generate the data for this situation and estimate the average yield (in kg) per tree of guava of the village along with its standard error.

With this we have come to the end of the unit. Let us summarise what we have discussed in this unit.

14.4 SUMMARY

In this unit, we have discussed the following points.

1. A number of examples are given illustrating the basic principles of cluster and multistage sampling. Also, we talked about some advantages that these sampling techniques have.
2. For equal and unequal size of clusters, some of the relations used in estimating the population are discussed. A number of examples are discussed to illustrate their use. Also, relative efficiency (RE) aspect of cluster sampling is discussed with help of some examples.
3. Examples are discussed to illustrate the use of the following formulations for estimating population mean in case of two-stage sampling method.

$$(a) \hat{Y}_{2s} = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m y_{ij} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i$$

$$(b) V(\hat{Y}_{2s}) = \left(\frac{1}{n} - \frac{1}{N} \right) S_b^2 + \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M} \right) S_w^2, \text{ where}$$

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^N (\bar{Y}_i - \bar{Y})^2, \text{ and}$$

$$S_w^2 = \frac{1}{N(M-1)} \sum_{i=1}^N \sum_{j=1}^M (Y_{ij} - \bar{Y}_i)^2.$$

$$(c) \hat{V}(\hat{Y}_{2s}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_b^2 + \frac{1}{N} \left(\frac{1}{m} - \frac{1}{M} \right) s_w^2 \\ = \left(\frac{N-n}{Nn} \right) s_b^2 + \frac{1}{N} \left(\frac{M-m}{mM} \right) s_w^2, \text{ where}$$

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_i - \hat{Y}_{2s})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n \bar{y}_i^2 - n \hat{Y}_{2s}^2 \right], \text{ and}$$

$$s_w^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \bar{y}_i)^2 \\ = \frac{1}{n(m-1)} \left[\sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 - m \sum_{i=1}^n \bar{y}_i^2 \right].$$

14.5 ANSWERS/SOLUTIONS

E1) Of course, this you will have to do it yourself.

E2) Change the figures and proceed as in Problem 3.

E3) (*Hint*) Proceed as in Problem 3 to find the values of \hat{Y}_c and $\hat{V}(\hat{Y}_c)$. Finally, proceed as in Problem 4 to calculate $\hat{V}(\hat{Y})$. Hence, these values will give

$$RE = \frac{\hat{V}(\hat{Y})}{\hat{V}(\hat{Y}_c)}.$$

E4) Do it yourself, using relations given in Table 3.

E5) Let the generated data be as in the following table.

Cluster	M	m	ssu's		y _i	\bar{y}_i
1	4	2	5	15	20	10
2	4	2	1	7	8	4
3	4	2	26	10	36	18
4	4	2	7	15	22	11
5	4	2	22	6	28	14

Here, $N = 120$, $n = 5$, $M = 4$, and $m = 2$. Also, the sample means for selected cluster (psu) are as given in the last column of the table. Now, the estimate of average number of daily calls can be computed using estimator \hat{Y}_{2s} . Using values from the last column of above table, we get

$$\hat{Y}_{2s} = \frac{1}{5} \sum_{i=1}^5 \bar{y}_i = \frac{1}{5} [10 + 4 + 18 + 11 + 14] = 11.4.$$

Also, we know that the estimated variance of \hat{Y}_{2s} is given by the relation

$$\hat{V}(\hat{Y}_{2s}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_b^2 + \frac{1}{N} \left(\frac{1}{m} - \frac{1}{M} \right) s_w^2.$$

Here, using values from table and the value of \hat{Y}_{2s} , we get

$$\begin{aligned} s_b^2 &= \frac{1}{n-1} \left[\sum_{i=1}^n \bar{y}_i^2 - n\hat{Y}_{2s}^2 \right] \\ &= \frac{1}{5-1} [(10)^2 + (4)^2 + \dots + (14)^2 - 5(11.4)^2] \\ &= \frac{1}{4} [757 - 649.8] = 26.8. \end{aligned}$$

Next, to calculate s_w^2 , we first compute the term involving sum of squares of all the individual observations. Thus, using values from above table, we have

$$\sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 = (26)^2 + (34)^2 + \dots + (29)^2 = 9446, \text{ and}$$

$$\sum_{i=1}^n \bar{y}_i^2 = (30)^2 + (36)^2 + \dots + (26)^2 = 4554.$$

Thus,

$$\begin{aligned} s_w^2 &= \frac{1}{n(m-1)} \left[\sum_{i=1}^n \sum_{j=1}^m y_{ij}^2 - m \sum_{i=1}^n \bar{y}_i^2 \right] \\ &= \frac{1}{5(2-1)} [9446 - 2 \times 4554] \\ &= \frac{1}{5} [338] = 67.6. \end{aligned}$$

Thus, the estimate of variance is

$$\begin{aligned} \hat{V}(\hat{Y}_{2s}) &= \left(\frac{N-n}{N} \right) s_b^2 + \frac{1}{N} \left(\frac{M-m}{M} \right) s_w^2 \\ &= \left(\frac{25-5}{25 \times 5} \right) (26.8) + \frac{1}{25} \left(\frac{4-2}{4 \times 2} \right) (67.6) \\ &= \left(\frac{20}{125} \right) (26.8) + \left(\frac{2}{200} \right) (67.6) = 4.29 + 0.676 = 4.97. \end{aligned}$$