UTTAR PRADESH
RAJARSHI TANDON OPEN UNIVERSITY

UGMM - 11
Probability and Statistics

Block

1

DESCRIPTIVE STATISTICS

Course Design Committee

Prof. S.K. Mitra *(Chairman)*
Indian Statistical Institute
New Delhi

Prof. A.M. Goon
Presidency College
Calcutta

Prof. J. Medhi
Guwahati

Prof. B.L.S. Prakasa Rao
Indian Statistical Institute
New Delhi

Prof. Aloke Dey
Indian Statistical Institute
New Delhi

Prof. K. Balasubramanian
Indian Statistical Institute
New Delhi

Prof. D.D. Joshi
Ex-Pro-Vice-Chancellor
IGNOU

Dr. V. Madan
School of Sciences
IGNOU

Dr. Poornima Mital
School of Sciences
IGNOU

Dr. Manik Patwardhan
School of Sciences
IGNOU

Dr. Sujatha Varma
School of Sciences
IGNOU

Block Preparation Team

Prof. S.K. Mitra *(Editor)*
ISI, New Delhi.

Prof. Aloke Dey *(Co-editor)*
ISI, New Delhi

Prof. A.M. Goon
Presidency College
Calcutta

Prof. G.S. Rao *(Language Editor)*
IGNOU

Dr. Manik Patwardhan
School of Sciences
IGNOU

Course Coordinator : Dr. Manik Patwardhan

Production

Mr. Balakrishna Selvaraj
Registrar (PPD)
IGNOU

Mr. M.P. Sharma
Joint Registrar (PPD)
IGNOU

Acknowledgement

To Dr. Parvin Sinclair and Dr. Sujatha Varma for their useful comments on the manuscript.

December . 1992

ISBN-81-7263-199-5

PROBABILITY AND STATISTICS

Inflation up by 12.8%.
The wife of a smoker is four times more likely to get cancer than that of a non-smoker.
Mosquito popuiation peaks in April.

All these headlines appeared in a daily newspaper during April 1992. You, too, must have often come across these or similar news items. These conclusions have one thing in common. They have all been arrived at with the help of statistics. Statistics is a body of concepts and methods used to collect and interpret data. It is used to draw conclusions in situations where uncertainty prevails. Actually, in our everyday life, each one of us often analyses data and draws conclusions, although unconsciously. For example, we are sure you have a favourite shop where you buy vegetables. How did you zero in on that shop? You must have experimented with many other shops, and gauged the quality and the freshness of the vegetables sold there before opting for one particular shop. In this course, we will acquaint you with some methods which ensure that your choice is the right one!

Whenever conclusions are drawn after analysis of data, their credibility depends on the methods used and the care exercised in the data collection. In the first block of this course, we shall talk about collection and organisation of data. We shall also discuss various descriptive measures of data like the measures of central tendency, dispersion, skewness and kurtosis. We shall also, briefly, talk about the organisation of bivariate data before discussing correlation and regression.

After collecting and organising data, the next important task is to analyse it. Probability theory plays an important role in the analysis and interpretation of statistical data. We discuss this theory in Block 2. After introducing the basic concepts of probability, we'll acquaint you with some standard discrete frequency distributions like Bernoulli, binomial, hypergeometric and Poisson.

In Block 3, we continue our discussion of probability distributions. But this time we deal with the continuous ones. Here also we'll discuss some standard continuous distributions like uniform, exponential and normal. We then go on to discuss bivariate distributions. You will also study the χ^2, t and F distributions. Next, we state some important limit theorems: weak law of large numbers, central limit theorem.

The last block, Block 4, deals with statistical inference. It concerns drawing conclusions about a population on the basis of the analysis of sample data. In this block, we shall introduce you to the techniques which ensure that the conclusions drawn are the best under the given circumstances.

Throughout this course, we have tried to explain concepts with the help of numerous examples. Studying these examples will help in understanding the concepts. We have also given exercises related to each concept that we discuss. Do solve them as and when you encounter them. You can tally your answers with ours, given at the end of the units. It would help if you have a calculator to do the computations in the exercises. You may decide to rely on your brain power and not use a calculator. But that will be time-consuming. Calculators are also available at your study centre. If you need further information about the concepts in this course, you may consult the following books :

1) *An Outline of Statistical Theory*, Vols. I & II by Goon, Gupta and Dasgupta,
2) *An Introduction to Probability Theory and Mathematical Statistics* by V.K. Rohatgi
3) *Introduction to Mathematical Statistics* by R.V. Hogg and A.T. Craig.

These are available in your study centre library.

BLOCK 1 DESCRIPTIVE STATISTICS

The word "Statistics" was first used to describe data collected by the government. Even today you know that our government collects data on various subjects. Recently (in 1991, to be exact), we have had our tenth census of human population. The data collected in the census are used to formulate economic policies as well as those on public health, education, trade and commerce, etc. Apart from the government, there are many other agencies which collect data. For example, a soap manufacturer may collect data about population preferences before launching a new brand of soap. IGNOU may collect data to find if there is a need for a particular programme before launching it. In this first block, we shall see how data are collected and organised.

In the first unit, we acquaint you with various statistical terms which we'll be using in this course. In this unit, we also talk about the collection of raw data. We'll see how raw data can be put in the form of a frequency distribution. We shall also look at some ways of diagrammatic representation of the data.

In the second unit, we talk about two kinds of measures : measures of central tendency and those of dispersion. These measures summarise the data and yet give us a fair amount of information about it.

In Unit 3, we take up the study of skewness and kurtosis. A measure of skewness tells us whether a given frequency distribution is symmetric or not. It also indicates the degree of asymmetry. On the other hand, a measure of kurtosis gives us the degree of flatness of the frequency distribution.

We then extend the techniques of collection and organisation of data discussed in Unit 1 to apply to bivariate data in Unit 4. Here, we also develop methods which tell us whether the two variables show a linear relationship or not. You will also see how to fit a straight line to given bivariate data. Once such a line is fitted, the values of the dependent variable can be predicted for specific values of the independent one with the help of the fitted line.

To illustrate the concepts occurring in Units 2, 3 and 4, we will refer, again and again, to the data sets in Unit 1. So make sure that you are familiar with them. And the best way to ensure that you have understood any concept is to solve the exercises related to it. So, good luck!

Notations and Symbols

x_i	Class mark of the i^{th} class
f_i	frequency of the i^{th} class
$\bar{x}$	mean
$\tilde{x}$	median
$\overset{\circ}{x}$	mode
q_i	i^{th} quartile
v	coefficient of variation
Var(x)	Variance of x
s	standard deviation about $\bar{x}$
s_A	root mean square deviation about A
MD_A	mean absolute deviation about A
R	range
m_r	r-th moment about $\bar{x}$
m'_r	r-th moment about zero
$m'_r(A)$	r-th moment about A
Sk_i	measure of skewness (i=1,2,3,4)
b_2	measure of kurtosis
z_p	p-quantile
f_{ij}	frequency of the (i, j)–th cell
b_{yx}	regression coefficient of y on x
Cov (x,y)	covariance of x and y
τ	correlation coefficient

Greek Alphabets

α	Alpha
β	Beta
γ	Gamma
δ	Delta
ε	Epsilon
ζ	Zeta
η	Eta
θ	Theta
ι	Iota
κ	Kappa
λ	Lambda
μ	Mu
ν	Nu
ξ	Xi
o	Omicron
π, Π	Pi (capital pi)
ρ	Rho
σ, Σ	Sigma (capital sigma)
τ	Tau
υ	Upsilon
ϕ	Phi
χ	Chi
ψ	Psi
ω	Omega

UNIT 1 FREQUENCY DISTRIBUTION OF A CHARACTER

Structure

1.1 INTRODUCTION

In this unit, we shall talk about the basics of statistics. We shall define the terms which we shall be using again and again throughout this course. It is possible that you have read all this before. But that might have been some years ago. So a quick look through this unit will help you to recall the relevant facts. In case you have never been introduced to statistics before, this unit will gradually acquaint you with its basic concepts. You will find that most of the terms we use in statistics are part of our daily vocabulary. But we have to know their precise meaning before we use them in statistics.

Further, you will see how to collect the data relating to a given investigation. You will also be introduced to the concept of frequency distributions. Through simple examples, we shall acquaint you with the various modes of presenting a frequency distribution—tabular as well as diagrammatic.

Objectives

On reading this unit, you should be able to :

- distinguish between a qualitative and a quantitative character,
- differentiate between a discrete and a continuous variable,
- draw up a frequency table and get the relative frequencies, cumulative frequencies and frequency densities,
- decide upon a suitable mode of representing a frequency distribution diagrammatically.

1.2 RAW MATERIALS OF STATISTICS

We have told you that in this unit we are going to define some basic terms which occur frequently in statistics. How about starting with the word "statistics"? We use the term "statistics" in two different contexts. **Numerical data** arising in some sphere of life, as well as the **discipline** that concerns itself with the collection, analysis and interpretation of such data are both called statistics.

'data' is the plural of 'datum'.

For example, we talk about

- the admission statistics of IGNOU,
- the statistics of steel production in India, or
- the statistics of the Indian team's performance in international cricket tests.

In all these cases, we are talking about numerical data.

On the other hand, when we talk about

- a student of statistics or
- a book of statistics,

We have the discipline in mind.

Now let us turn our attention to the two concepts of "character" and "individual" which are basic to any statistical study. To understand these two terms, we consider the following cases :

1) A teacher looks at the grades (say A,B,C,D and E) awarded to his students on their performance in an examination. Here, the students are the individuals and the character is the grade (per student).

2) An economist collects data on the size and the expenditure on food in a given month for urban households. In this case, the individuals are the households. What about the character? Here we see that there are two characters under study, namely, household size and expenditure on food (per household) in the given month.

Thus, in any instance, the data relate to one or more **characters** and a group of **individuals** who possess the character or characters in varying forms or amounts.

Further, we can classify the data as primary or secondary. If we collect our own data on the relevant group of individuals and use it in a study, then the data will be called **primary.** In some cases, however, we may choose to make use of the data already available in government publications or the data collected by some other agency. Such data are said to be **secondary**. We can save a lot of time and money if we use secondary data. But, at the same time, we have to be very careful. We have to make sure that

- the data are relevant to our enquiry,
- the concepts and definitions used conform to what we have in mind, and
- the data are reliable.

On the other hand, if we decide to use primary data, we shall have to decide on how to go about collecting it. Primary data can be obtained in a number of ways, depending on the information sought and also on our knowledge of the relevant group of individuals. We give below some of the commonly used methods.

1) Direct Observation

Suppose we want to know the number of leaves per twig of a tree, or the weight (in grams) per egg in a basket of eggs or the health status (good/indifferent/poor) per student in a class. In each of these cases, we can obtain the required information by direct observation, through counting or measurement or, simply, by inspection.

But in social and behavioural sciences, we collect information from persons who are supposed to know. These persons are called **informants.** We can either get the information directly from the informants or through intermediaries (called **enumerators**) appointed for the purpose. In such cases, we can use the following methods.

2) Questionnaire Method

If the informants happen to be sufficiently enlightened, then we can give them blank **questionnaire** forms and request them to provide the necessary information by filling out the forms. This method would be appropriate in gathering information about say, the attitude of doctors towards euthanasia (mercy-killing).

3) Interview Method

In case the informants are illiterate or not enlightened enough, the enumerators fill out the **schedule** by a thorough and tactful questioning of each informant. As you are aware, this method is used in the population census held once in ten years in our country.

Now solve these exercises and check whether you have grasped these ideas or not.

E 1) Indicate which of the following are primary data and which are secondary :

a) Data taken from the Government of India publication, *Statistical Abstract India* of 1986.

b) Data collected by a market research bureau through door-to-door enquiry to study the demand for a newly marketed shaving lotion.

c) Data collected by a medical research group through questioning of patients visiting a hospital's out-door facilities.

d) Weather data recorded by the Department of Meteorology and then used by the investigator for writing a Ph.D. thesis.

E 2) What mode of data collection would you recommend for

a) studying the progress of a public health programme covering a city's slums?

b) finding out the reactions of a number of economists to this year's budget proposals?

c) estimating the yield rate (per acre) of a particular variety of wheat?

d) estimating the time taken to complete a particular calculation?

We have observed before that data relate to one or more "characters". Let us look at this term more closely.

Characters fall into two broad categories.

There are certain characters which take varying forms for different individuals but cannot be expressed numerically. The brand name of motor cars plying in an Indian city is such a character; it may be Ambassador Contessa, Premier Padmini Deluxe, Standard Herald Gazelle, Maruti 1000 or other. The employees in a city hospital may be observed for their smoking habits; any given employee will then be recorded as a smoker or a non-smoker. Such a character, whose possible forms can be distinguished verbally, but not numerically, is called a **qualitative character** (or **attribute**).

On the other hand, we can express characters like the size of families, age of teachers, height of students, weight of eggs, etc., in numerical or quantitative terms. The size of a family (i.e., the number of members in the family) will be a positive integer—1,2,3, etc. The age of a teacher may be given in years or in years and months. The height of a student may be given in centimetres and may be rounded off to the nearest centimetre. The weight of an egg may be recorded in grams and again may be rounded off to the nearest tenth of a gram. Such characters are called **quantitative characters** (or **variables**).

A qualitative character, too ultimately yields numerical data. This is because we will finally note how many of the individuals under study have any given form of the character. In the case of motor cars in a city, we thus note how many of the cars are Ambassador Contessas, how many are Maruti 1000, and so on. However, the data on a quantitative character are numerical right from the beginning and so we can give them a more in-depth statistical treatment than those on a qualitative character. A qualitative character whose forms have an implied ranking (or gradation), however, stands on a somewhat different footing. We can assign scores to these forms and thus, express the raw data in quantitative terms. Data of this type are called **ordinal data.** For example, an employee's performance in a year may be very good, good, satisfactory, bad or poor. But we can assign the scores 5,4,3,2 and 1, to these five categories, and immediately the data on the performance of the employees in an office assume a numerical look. Surely, there is a lot of arbitrariness in assigning scores this way. Nevertheless, this method of scoring is quite popular with research workers in social and behavioural sciences.

Note that 'scoring' must be distinguished from 'coding' used to facilitate the processing of data on an electronic computer. We use codes mainly for identification purposes, similar to the use of roll numbers in IGNOU. Scores, on the other hand, are more informative. For example, if you get a B grade in MTE-11, it means that you have a good grasp of the course.

See if you can distinguish between variables and attributes now.

E 3) Classify the following characters as qualitative or quantitative.

a) word-length (i.e., number of letters per word) of the words of a poem;

b) diameter of balls (in cm) produced by a firm;

c) mother tongue of the residents of a city;

d) attitude towards family planning of the couples living in a locality;

e) proportion of males in each group of 25 students.

We have classified characters into two categories: qualitative and quantitative. Now quantitative characters or variables, in their turn, may be classified as **discrete** and **continuous.**

A discrete variable is one that can conceivably assume only some discrete, or isolated values. The size of families, the proportion or the number of males in each group of 25 students, or the length of a word are variables of this type. The size of a family or the length of a word may take values like 1,2,3, etc., but no values in between. The number of males in a group of 25 students may be 0,1,2,...,24 or 25, while the proportion of males may be 0,0.04,0.08....,0.96 or 1; values in between these numbers are inconceivable.

A continuous variable, on the other hand, can possibly take any value in some interval. For example, the age (in years) of teachers, the height (in cm.) of students, the weight (in grams) of eggs are all continuous variables. Supposing the minimum age at which a person can join the teaching profession is α years and that every member of the teaching community has to retire on reaching the age β years, then the age of teachers must vary between α and β and can take any value within the interval $[\alpha, \beta]$. Indeed, the actual age of a teacher may well be 32.119237 years! However, there will be hardly any need to record the age with this much precision! The enquirer may be satisfied by taking the age correct to the second decimal place so that the teachers age may be recorded as 32.12 years. This is an example of how limitations of the measuring instruments can introduce a discreteness into the observations of a continuous variable. Similarly, the actual monthly income of an Indian which is a continuous variable, has to be expressed in rupees or in rupees and paise, since the paisa happens to be the smallest denomination coin in the Indian system of currency. This is also the case with the score in an examination of students taking the examination. The score is invariably expressed in integers and yet it has to be regarded as a continuous variable. This is because the score is supposed to measure the proficiency of the students in the subject concerned, and the proficiency may be taken to vary in a continuous manner (say, between 0 and 100).

Try this exercise now.

E 4) Indicate which of the following variables are discrete and which are continuous :

a) diameter of ball-bearings produced by a steel mill;

b) number of beds per hospital in a city;

c) proportion of heads in sets of 10 tosses of a coin;

d) length (in mm) of needles produced by a factory;

e) weight of loaves (in kg) produced by a bakery;

f) size of households in a village.

The distinction between a discrete and a continuous variable is important. Quite often, the statistical analysis of the data will differ accordingly. In fact, there are some techniques of statistical inferenee, which are based on the assumption that the variable under study is continuous. These are clearly inapplicable to data on a discrete variable.

In the next section, we shall discuss the concept of frequency distributions of qualitative characters and variables.

1.3 FREQUENCY DISTRIBUTIONS

In this section, we shall discuss the method of organising raw data into frequency distributions. You will see that we can get information out of a frequency distribution more easily than out of raw data. Here, we shall first discuss ungrouped frequency distributions and then discuss grouped ones.

1.3.1 Ungrouped Frequency Distributions

We use ungrouped frequency distributions when the data is of a qualitative nature, or when the variable under consideration is discrete. Here, we will take one example of each situation for illustration.

Frequency Distribution of a Qualitative Character

A botanist obtained a variety of linseed by cross-breeding of two pure varieties. She observed the colour of flowers of plants grown through inbreeding of the new mixed type (called plants of the F_2 generation). On the basis of these observations, she prepared the following table.

Table 1 : Classification of flowers in an F_2 population of linseed by colour

Colour	Number of flowers (frequency)	Relative frequency
Blue	169	0.538
Lilac	61	0.194
White	62	0.197
Pink	22	0.070
Total	314	0.999

(Ref: *Statistical Methods for Agricultural Workers* by Panse and Sukhatme).

The figures in the second column of Table 1 are called the **frequencies** of the four classes (or of the four colours). So 'frequency' indicates how frequently the corresponding form of the character under study (viz., colour) occurs in the collected data. The sum of the frequencies, 314 in this case, is said to be the **total frequency**. The first two columns in Table 1 constitute a **frequency table.** Since these indicate the manner in which the total frequency 314 (or the total number of individuals) is distributed among the four classes, they are also said to represent the **frequency distribution** of colour for the 314 flowers. Perhaps a better expression is 'the frequency distribution of the 314 flowers by colour'.

Alternatively, we can also write the frequency distribution in terms of the proportions of blue, lilac, white and pink flowers in the group. These proportions give the **relative frequencies,** and are shown in the third column of Table 1. By definition,

$$\text{relative frequency of a class} = \frac{\text{frequency of the class}}{\text{total of frequency}} \quad \ldots (1)$$

Then what is the total relative frequency? One, of course. But you can see that in Table 1, the relative frequencies do not add up exactly to 1. This is because the individual figures are all approximate, rounded off to a certain number of decimal places.

Note that while the distribution of frequencies answers questions of the type 'How many flowers in the given group are blue?', the relative frequency has to do with questions like 'what is the proportion (or percentage) of blue flowers in the group?'

Further, in any situation, a frequency must be non-negative integer. The value 0 is admissible, for in the above situation it is conceivable that we might have a fifth flower colour, say yellow, which was absent in the sample. A relative frequency, on the other hand, must be a rational number in the interval [0,1].

The simplest type of classification of a group of individuals by a qualitative character is a **dichotomy**, i.e., a classification with just two classes. A group of students may thus be classified by sex as boys and girls or by performance at an examination as successful and unsuccessful.

Fig. 1 : Early notch-cutting by primitive man

Let us now take an example of the data on a discrete variable.

Ungrouped Frequency Distribution of a Discrete Variable

Consider the data collected by a social scientist on household size for households in an urban locality, given in Table 2.

Table 2 : Data on household size for 80 households in an urban locality

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

As in the case of a qualitative character, here too, we would like to summarise the data by forming a frequency table. For this it would be necessary to count the number of times 1 appears, the number of times 2 appears and so on. We can count more easily if we follow a tallying system. This system can be used by people without any formal training in arithmetic (like our cave-dwelling forebears!)

Thus, we take nine classes defined by the nine distinct values 1,2,...9, noting that 9 was the largest household size recorded in the data. The second column in Table 3 shows the tallies against each of these values. After counting the tallies, we write the frequencies in the third column. In the fourth column we have written the relative frequencies.

Table 3 : Frequency table for size of 80 households

Household size	Tallies	Frequency	Relative frequency
1	|||	3	0.0375
2	𝍸 ||	7	0.0875
3	𝍸 𝍸 |	11	0.1375
4	𝍸 𝍸 ||||	14	0.1750
5	𝍸 𝍸 𝍸 ||||	19	0.2375
6	𝍸 𝍸 ||	12	0.1500
7	𝍸 |||	8	0.1000
8	||||	4	0.0500
9	||	2	0.0250
Total		80	1.0000

There are two more ways in which we can represent the frequency distribution of discrete variable. Both make use of what are called the cumulative frequencies of the variable. For a discrete variable like household size, the frequencies answer questions of the type : 'How many individuals in the given group have the value k of the variable?', and the relative frequencies answer questions like : 'What proportion o the individuals has the value k of the variable?' But how do we answer a question like "How many individuals have the value k or less?" or "How many individuals have the value k or more?"

From Table 3, you can see that the number of households of size k or less is 3 for k=1, 3+7=10 for k=2, 10+11=21 for k=3, and so on. We obtained these figures by taking cumulative totals of the frequencies in Table 3, starting from the lowest observed value of the variable and going successively to the higher values. These are called **cumulative frequencies** of the **less than type.** Similarly, to get the number of

households having size k or more, we take the cumulative total of the frequencies in Table 3, starting from the highest observed value of the variable and moving successively to the lower values. The figures obtained in this manner are called **cumulative frequencies** of the **more than type**. Cumulative frequencies of the more than type provide one mode of representation of the frequency distribution of a variable; those of the less than type provide another. We illustrate these two modes for the data on household size by means of Tables 4a and 4b.

We cannot talk of cumulative frequencies of a qualitative character unless it is of the ordinal type.

Table 4a : Cumulative frequency table of "less than" type for size of 80 households.

Household size	Cumulative frequencies
$\leqslant 1$	3
$\leqslant 2$	10
$\leqslant 3$	21
$\leqslant 4$	35
$\leqslant 5$	54
$\leqslant 6$	66
$\leqslant 7$	74
$\leqslant 8$	78
$\leqslant 9$	80

Table 4b : Cumulative frequency table of "more than" type for size of 80 households

Household size	Cumulative frequencies
$\geqslant 1$	80
$\geqslant 2$	77
$\geqslant 3$	70
$\geqslant 4$	59
$\geqslant 5$	45
$\geqslant 6$	26
$\geqslant 7$	14
$\geqslant 8$	6
$\geqslant 9$	2
any value greater than 9	0

While making use of Table 4a, you should remember that the cumulative frequency of the less-than type is 0 for any value of the variable less than 1, is 3 for any value between 1 and 2 but less than 2, is 10 for any value between 2 and 3 but less than 3, and so on. Finally, the cumulative frequency of the less than type is 80 (the total frequency) for 9 or any value exceeding 9.

Similarly, the cumulative frequency of the more-than type is 0 for any value of the variable exceeding 9, is 2 for any value between 8 and 9 but exceeding 8, is 6 for any value between 7 and 8 but exceeding 7, and so on. Finally, the cumulative frequency of the more than type is 80 for the value 1 or any value less than 1.

Thus, we can see that the cumulative frequencies are constant in some intervals, but when they change, they change in jumps.

It goes without saying that by taking cumulative total of the relative frequencies (or by dividing the cumulative frequencies by the total frequency), we can form two other tables : a table of cumulative proportions of the less than type and a table of cumulative proportions of the more-than type. The former would provide answers to questions like, 'What is the proportion of individuals having the value of the variable less than or equal to k?' The latter would answer questions like, 'What is the proportion of individuals having the value of the variable greater than or equal to k?

If you have understood the discussion so far, you will surely be able to do these exercises.

E 5) The following is a record of the results of an opinion poll conducted among the 55 inmates of a nursing home to know their assessment of the services offered by the home:

G	B	V	P	B	G	B	S	G	S	S
G	S	G	V	B	B	G	S	V	B	G
S	G	B	B	G	G	V	G	G	S	B
V	S	S	G	S	V	S	B	S	G	S
B	V	S	S	B	S	S	S	B	G	G

[V=very good, G= good, S=satisfactory, B=bad, P=poor]

Draw up a frequency table and a relative frequency table for these data. Hence, answer the following questions:

a) How many of the inmates think the services are good?

b) How many think that the services are at least satisfactory?

c) What is the percentage of inmates who consider the services to be less than satisfactory?

E 6) The following data indicate the length per word for the 91 words in Tagore's poem 'Where the mind is without fear and the head is held high, etc.':

5	4	3	5	8	6	6	3	4	5
3	4	4	5	8	2	6	7	4	
4	5	6	4	9	6	4	2	6	
2	9	2	3	3	3	2	4	2	
7	2	4	4	4	3	4	4	7	
4	4	9	3	7	4	5	4	2	
3	5	2	5	10	3	5	8	6	
3	3	6	2	5	3	3	7	3	
4	5	8	5	3	4	4	3	2	
2	3	5	5	5	3	2	6	7	

Draw up a frequency table. In the same table, show the relative frequencies and the cumulative frequencies of both types. Hence, answer the following questions :

a) How many of the words have at least 6 letters?
b) How many have 5 letters or more?
c) What is the proportion of words with 2 letters?
d) What is the proportion of words of length 4 or more?

Until now, we have seen how to construct ungrouped frequency tables for qualitative characters and discrete variables. For such a table, we count the frequencies of each distinct attribute or value taken by the variable, and so there is no loss of information. But this may not always be feasible. For example, suppose we have raw data on the number of grains per earhead for 400 ears of a variety of wheat. It is quite possible that there are some earheads with as few as 8 grains and some with as many as 57 grains. In this case, if we construct a frequency table taking each distinct value between 8 and 57, then the table would be too long. Then again, ungrouped frequency tables cannot be constructed for data on continuous variables, because a continuous variable can take infinitely many distinct values. In such cases, then, it becomes necessary to group some variable values together and then construct frequency tables. We shall discuss such grouped frequency distributions in the next sub-section.

1.3.2 Grouped Frequency Distribution

To illustrate the method of construction of a grouped frequency table, we consider the data collected by a botanist in Shillong, shown in Table 5. Note that we are dealing with a continuous variable here.

Table 5 : Petiole length (in cm.) of 198 leaves of a four-year old pipal tree (see Fig. 2)

4.5	5.4	5.3	6.3	5.7	5.5	4.1	2.9	2.7	6.0	5.9	1.8	3.7	4.1	5.6
2.6	3.0	6.0	7.8	4.5	5.7	4.5	8.0	5.5	7.5	3.1	3.1	5.2	6.8	9.2
5.5	4.5	5.5	7.0	4.5	4.0	5.9	3.8	6.0	5.2	5.6	7.0	6.3	5.1	6.0
6.3	4.5	5.0	5.3	5.6	6.3	3.4	5.1	6.7	6.2	7.2	6.2	5.0	6.1	6.3
4.7	4.1	6.1	5.6	5.5	4.4	6.0	5.0	3.4	5.0	2.5	5.7	5.2	6.1	6.5
5.6	5.5	4.5	5.5	7.7	7.0	7.3	6.5	6.7	6.1	6.7	4.7	8.5	4.7	6.7
4.6	8.2	6.9	3.9	7.2	4.2	6.1	1.6	7.2	6.5	3.6	5.9	5.3	6.6	5.0
4.2	1.9	2.2	5.2	6.6	4.9	5.9	5.4	6.5	6.6	6.8	4.1	4.7	5.7	4.1
5.7	5.0	5.7	5.2	2.8	4.3	4.6	4.9	6.0	5.9	4.5	3.7	5.7	3.8	5.6
5.2	3.9	6.5	5.0	5.2	6.0	2.3	5.2	3.2	5.5	7.1	7.0	3.2	7.2	5.9
5.3	1.6	6.9	6.1	6.3	6.7	2.4	6.3	4.8	4.6	6.7	1.5	6.8	5.9	5.3
7.0	4.3	6.7	5.4	4.7	5.1	5.2	7.4	4.5	6.4	5.0	2.0	5.7	4.6	4.9
5.2	6.0	4.5	6.1	3.5	5.9	5.0	6.8	5.0	1.0	5.5	4.9	5.9	5.2	6.1
0.8	5.3	5.9												

Here, the values are recorded correct to one decimal place (i.e., correct to the nearest tenth of a centimetre). The lowest observation in the set is 0.8 and the highest 9.2. If we take our classes as 0.6-1.3, 1.4-2.1, 2.2-2.9,....,8.6-9.3, then the total number of classes will be 11. To get the frequencies for these classes, we again go in for the tallying system which we had adopted in Table 3. This is done in Table 6.

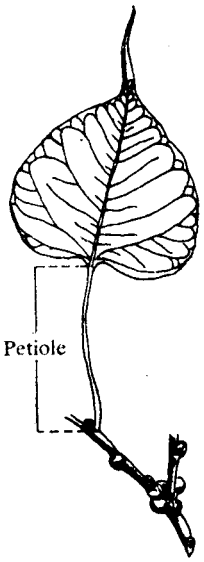


Fig. 2 : Petiole length

Table 6 : Frequency table for the data of Table 5 on petiole length of leaves of a pipal tree

Petiole length (cm)	Tallies	Frequency
0.6 – 1.3	||	2
1.4 – 2.1	𝍸 |	6
2.2 – 2.9	𝍸 |||	8
3.0 – 3.7	𝍸 𝍸	10
3.8 – 4.5	𝍸 𝍸 𝍸 𝍸 ||||	24
4.6 – 5.3	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	43
5.4 – 6.1	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||	52
6.2 – 6.9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	33
7.0 – 7.7	𝍸 𝍸 𝍸	15
7.8 – 8.5	||||	4
8.6 – 9.3	|	1
Total		198

But here the classes need to be redefined. The reason is that the value recorded as, say, 4.6 actually stands for some value between 4.55 and 4.65. Similarly the value 5.7 stands for some value between 5.65 and 5.75. Thus, the class taken as 4.6–5.7 in Table 6, in fact, begins at 4.55 and ends at 5.75. The other classes have to be viewed in the same way. We then have to properly define the classes in terms of class-intervals, with no gap between any two successive intervals. The two end-points of a class-interval are called **class boundaries** (the lower and the upper) while the mid-point is called the **class mark. The width** (or length) of a class interval is, of course, the difference between the upper class boundary and the lower. The end-values of a class, when the classes are defined as in Table 6, are called the **class limits** to distinguish them from the class boundaries. We may then say that the frequency table in the form of Table 7 presents the frequency distribution of petiole length more appropriately than does Table 6. The width of each class here is 0.8 cm.

Table 7 : Frequency/relative frequency table for petiole length of 198 leaves of a pipal tree

Petiole length (cm) Class interval	Frequency	Relative Frequency
0.55—1.35	2	0.0101
1.35—2.15	6	0.0303
2.15—2.95	8	0.0404
2.95—3.75	10	0.0505
3.75—4.55	24	0.1212
4.55—5.35	43	0.2172
5.35—6.15	52	0.2626

Table 7 continued

Petiole length (cm) Class interval	Frequency	Relative Frequency
6.15—6.95	33	0.1667
6.95—7.75	15	0.0758
7.75—8.55	4	0.0202
8.55—9.35	1	0.0050
Total	198	1.0000

The third column in Table 7 shows the same frequency distribution in terms of relative frequencies. While the frequencies tell us, for any interval, how many of the leaves have petiole length between the two class boundaries, the relative frequencies indicate what the proportion (or percentage) of such leaves is.

Here again, the cumulative frequencies of less than and more than types provide us with two additional modes of representing the same frequency distribution. You can see such representation in Table 8.

Table 8 : Cumulative frequencies for petiole length of 198 leaves of a pipal tree

Petiole length (cm) Class interval	Cumulative frequency	
	(less than type)	(more than type)
0.55—1.35	2	198
1.35—2.15	8	196
2.15—2.95	16	190
2.95—3.75	26	182
3.75—4.55	50	172
4.55—5.35	93	148
5.35—6.15	145	105
6.15—6.95	178	55
6.95—7.75	193	20
7.75—8.55	197	5
8.55—9.35	198	1

We have to be careful in interpreting the cumulative frequencies of either type for a continuous variable. In Table 8, 2 is the number of leaves having petiole length 1.35 cm or less, 8 is the number of leaves having petiole length 2.15 cm or less, and so on. Hence, the cumulative frequencies of **less than** type now correspond actually to the respective **upper class boundaries**. On the other hand, if we look at the column of cumulative frequencies of more than type in Table 8, then obviously the number of leaves having petiole length 8.55 cm or more is 1, the number of leaves with petiole length 7.75 cm or more is 5, and so on. Thus, the cumulative frequencies of **more than** type now correspond to the respective **lower class boundaries.**

But then there is yet another way of describing the frequency distribution of a continuous variable, viz., through the use of what are called the frequency densities of the different classes. By the **frequency density** of a class we mean the frequency per unit of width in the class. It is somewhat similar to the population density of a locality and is defined by the formula:

$$\text{frequency density of a class} = \frac{\text{class frequency}}{\text{class width}} \quad \ldots \quad (2)$$

The series of class intervals taken together with the series of frequency densities should give a good idea of the frequency distribution of the variable being studied.

You may wonder why we need to bring in frequency densities at all. Are'nt the frequencies supposed to give us an idea of the frequency distribution? But there are situations where we have to take classes of varying widths. In these cases, frequency densities become more meaningful. For example, consider the frequency distribution of a variable like monthly family income or family wealth at a given date. For any group of families, a large majority of the families will have incomes in the lower income brackets while the number of families will be smaller towards the higher

income brackets. As in other cases, here too, we may choose to have classes of the same width. However, if the common width is small, say Rs. 200, then too many classes will have to be taken. Many of these classes might be empty. This will bring in an irregular pattern and gross distortion in the true nature of the distribution. On the other hand, if the common width is large, say Rs. 1,000, then the number of classes will be too few and the true nature of the distribution, which usually shows rapid changes in the lower parts of the range, will get blurred. This will also lead to serious errors in the statistical measures computed on the basis of the grouped data. Therefore, it would be advisable to have classes of varying width—narrower classes in the lower parts of the income range and classes of increasing width towards the higher parts of the range. Now, when the classes are of varying width, the class frequencies will not be comparable. In such situations, the frequency densities that are obtained from the frequencies by reducing them to a common base (see Table 9) should be used.

Table 9 : Frequency distribution of monthly income for 1,276 urban families

Income (Rs)	Frequency	Frequency density
0	218	1.0900
200	153	0.7650
400	190	0.6333
700	152	0.5067
1000	159	0.3975
1400	119	0.2975
1800	107	0.2140
2300	73	0.1460
2800	49	0.0817
3400	23	0.0375
4000	15	0.0188
4800	8	0.0080
Total	1,276	—

Thus, frequency densities give us a true picture of the frequency distribution when the classes are of varying width. This will be all the more obvious when we consider the problem of diagrammatic representation of the frequency distribution of a continuous variable in the next section.

We have already mentioned at the end of Sec. 1.3.1 that even in the case of a discrete variable (or, for that matter, of a qualitative character), we may have to define the classes in terms of more than one distinct value of the variable (or more than one distinct form of the qualitative character). Table 10 illustrates this point. We would have to deal with as many as 50 classes if we did not use the type of condensation that is indicated by the first column of the table.

Table 10 : Frequency distribution of number of grains per earhead for 400 ears of a variety of wheat (see Fig. 3)

Number of grains per earhead	Frequency
8-12	1
13-17	17
18-22	25
23-27	86
28-32	125
33-37	77
38-42	55
43-47	9
48-52	4
53-57	1
Total	400

Source : *Statistical Methods for Agricultural Workers* by Panse and Sukhatme.

Fig. 3 : Wheat earhead

Now before we end this section, we list the main considerations guiding the construction of a frequency table.

For one thing, the classes should be **exhaustive**, in the sense that each of the observations should be assignable to one class or another.

Secondly, the classes should be **mutually exclusive.** This means that no two classes should overlap so that each of the observations can be assigned to exactly one of the classes without any ambiguity. These two criteria have clearly been followed in constructing the tables in Sections 1.3.1 and 1.3.2.

Thirdly, while it seems natural in the case of a qualitative character to take a separate class for each distinct form of the character and in the case of a discrete variable to take a class for each distinct value of the variable, the classes should not be too numerous. For the main objective is to summarise the data into an easily manageable and comprehensible form. Besides, having too many classes might result in a situation where many of the classes may have zero frequencies. Whereas the true distribution may show a gradual increase or decrease of frequency, the observed distribution, in such a case, will indicate abrupt changes in the frequency. Because of this, in many cases, we have to define the classes in terms of more than a single value of the character concerned.

But the classes should not be too few either. If there are too few classes, we are likely to overlook some important features of the distribution. For instance, an asymmetrical distribution may appear to be fairly symmetrical. We shall talk about symmetrical distributions in Section 1.4.4.

Further, in the computation of various measures related to the distribution, we assume that the observations within each class interval are concentrated at the class mark, instead of being spread over it. You will come across this in Unit 2. So, if the classes are too few, or equivalently, if each class is too wide, then this assumption may lead to considerable error in the computation of these measures.

Last, but not the least, we should see to it that in the case of a variable, the classes are defined in terms of the same number of distinct values of the variable or are of the same width. Otherwise, the frequencies (or the relative frequencies) for the different classes will not be comparable. On occasion we have to deviate from this rule, as you have seen from Table 9. In such cases, we have to work with frequency densities.

Try this exercise now.

E 7) Consider the data shown below :

Yield of seed cotton (in gm) for 120 plots of size 0.0005 acre

93	81	57	42	95	80	52	70	105	72	60	68
49	74	60	57	63	51	100	41	50	66	65	81
89	85	59	44	90	69	69	68	95	82	39	44
75	84	63	99	91	64	68	33	115	74	60	65
63	75	23	58	76	55	67	63	68	78	47	68
86	82	39	79	72	83	57	66	73	102	93	95
62	46	69	79	86	54	29	51	79	83	57	66
94	71	51	62	76	68	54	69	109	39	74	58
45	48	58	81	96	52	47	106	75	75	86	43
51	65	56	31	79	45	78	87	71	77	62	69

a) Draw up a frequency table with 10 classes. Also show, alongside the frequencies, the relative frequencies and the cumulative frequencies of both types.

b) Estimate the number of plots with an yield of
   i) 65.5 gm to 85.5 gm;
   ii) more than 100 gm; and
   iii) less than 60 gm.

c) What is the proportion of plots with yield
   i) between 70 gm and 100 gm ?
   ii) less than 75 gm ?
   iii) more than 105 gm ?

So far we have seen that a frequency distribution presents the data in a concise form. We can get a general idea of a distribution more readily and effectively through an appropriate diagram. In the next section, we talk about this diagrammatic representation.

1.4 DIAGRAMMATIC REPRESENTATION OF FREQUENCY DISTRIBUTIONS

We can use various kinds of diagrams to represent frequency distributions. In this section, we shall first see how to give a visual representation to the information in a frequency table. Then we shall talk about the representation of cumulative frequencies. After this, we shall discuss frequency curves, the diagrammatic representation of the frequency distribution of a variable which takes infinitely many values. Finally, we shall classify distributions into broad categories on the basis of their shapes. So let us start with the table of frequencies

1.4.1 Frequencies

We shall discuss the cases of ungrouped and grouped frequency distributions, one by one.

a) Case of an Ungrouped Frequency Distribution

An ungrouped frequency distribution of a qualitative character, given by the frequencies or the relative frequencies may be represented by means of what is called a **bar diagram**. The bars (actually rectangles) are as many as there are classes. These are taken perpendicular to the same base line, either vertically or horizontally. Further, the bars are equispaced and have the same width. Their height or length (as the case may be) indicates the frequencies (or relative frequencies) for the respective class. The frequency distribution for Table 1 is represented in the bar diagram in Fig. 4.

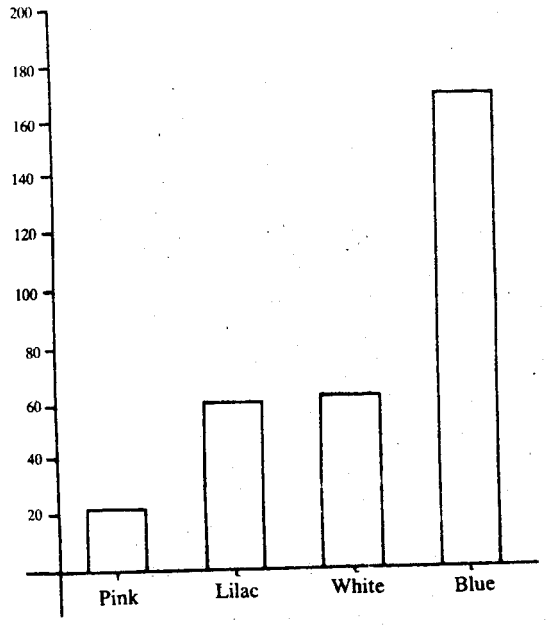


Fig. 4 : Bar diagram showing the frequency distribution of 314 flowers in an F_2 population of linseed by colour.

The relative frequencies can be represented by means of a bar diagram in the same way as the absolute frequencies. But in case of data on a qualitative character, a better mode of representing them would be to use what is called a **pie diagram** or **chart.** This diagram makes use of a circle, whose total area is divided into as many sectors as there are classes by drawing angles at the centre. The area of each section represents (is proportional to) the corresponding relative frequency. To illustrate the use of a pie diagram, let us consider the relative frequency table of colour of flowers in the F_2 generation of linseed (Table 1). We first determine the angles (in degrees) to be drawn at the centre of the circle (see Table 11).

Table 11 : Angles to be drawn at the centre of pie diagram for the frequency distribution of Table 1

Flower colour	Relative frequency	Angle to be taken
Blue	0.538	193.7°
Lilac	0.194	69.8°
White	0.198	71.3°
Pink	0.070	25.2°
Total	1.000	360.0°

The figures in the third column of the table indicate the measures (in degrees) of the angle to be drawn for each class, its sides extending from the centre of the circle to its circumference. Note that the angle for any given class measures

$360° \times$ relative frequency

Now, the area of a sector of angle θ radians in a circle of radius r is $\frac{1}{2}r^2\theta$. Thus, in a given circle, the area of a sector is directly proportional to its angle (whether in radians or in degrees). So if we draw sectors with angles given in the third column of Table 11, then the area of each sector is proportional to the corresponding angle which, in turn, is proportional to the corresponding relative frequency.

You can see the pie diagram corresponding to Table 11 in Fig. 5.

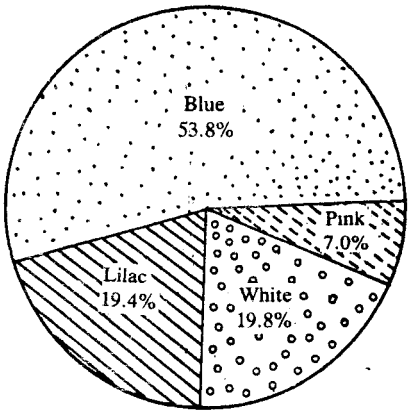


Fig. 5 : Pie diagram corresponding to Table 11.

If we are dealing with a discrete variable, we can also form a column diagram to represent its frequency distribution. In Fig. 6(a) we have the column diagram for the frequency distribution of household size (Table 3).

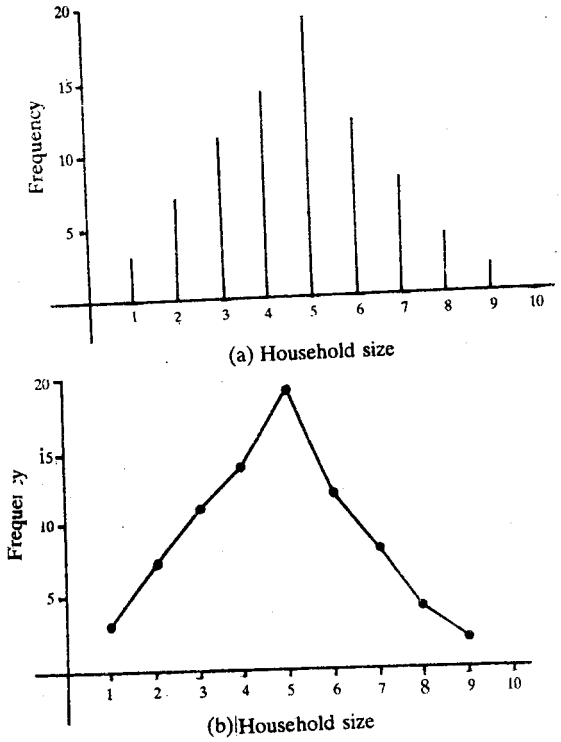


Fig. 6 : (a) Column diagram (b) Frequency polygon for the data on household size

When the possible distinct values of the variable are equispaced, as in the case of household size, word length, etc., an alternative mode of representing the frequency distribution is available to us. Again we take two mutually perpendicular axes, the horizontal for the variable and the vertical for the frequency (relative frequency). Then we plot each distinct value and the corresponding frequency (relative frequency) as a point on the graph paper, with respect to these axes. See Fig. 6(b) which represents the frequency distribution in Table 3. Then we join the points for the successive values of the variable by straight line segments. Next we take two additional points, one for the possible lower value than the lowest in the table and the other for the possible higher value than the highest in the table, the corresponding frequencies being of course, zero. For the distribution of household size, for instance, the two additional points will correspond to household size 0 and household size 10. Then we join these points with the points corresponding to the adjoining value, and thus obtain a closed polygon. Such a diagram is called a **frequency polygon**. Note that we can also get the frequency polygon by joining together the tops of the columns in the column diagram.

You may try your hand at drawing a frequency polygon now.

E8) Draw a column diagram and a frequency polygon to represent the frequency distribution of the data in E6.

---

b) Case of a Grouped Frequency Distribution

You would agree that the diagrammatic representation of a grouped frequency distribution has to be different from that of the ungrouped one. The reason for this is that unlike those for the ungrouped case, the frequencies in a grouped distribution are scattered over the different class intervals.

To represent the frequencies (relative frequencies), we again take two rectangular axes of coordinates, the horizontal for the variable value and the vertical for the frequency density (relative frequency density). Having marked the class boundaries on the horizontal axis, we draw on each class interval as base, a rectangle whose height equals the corresponding frequency density (relative frequency density). The area of each rectangle, therefore, represents the product of the class width and the frequency density (relative frequency density), i.e., the class frequency (relative frequency). The resulting diagram is called a **histogram**. In Fig. 7, we show you the histogram for the frequency distribution of petiole length per leaf of a pipal tree, drawn on the basis of the frequency densities given in Table 7.

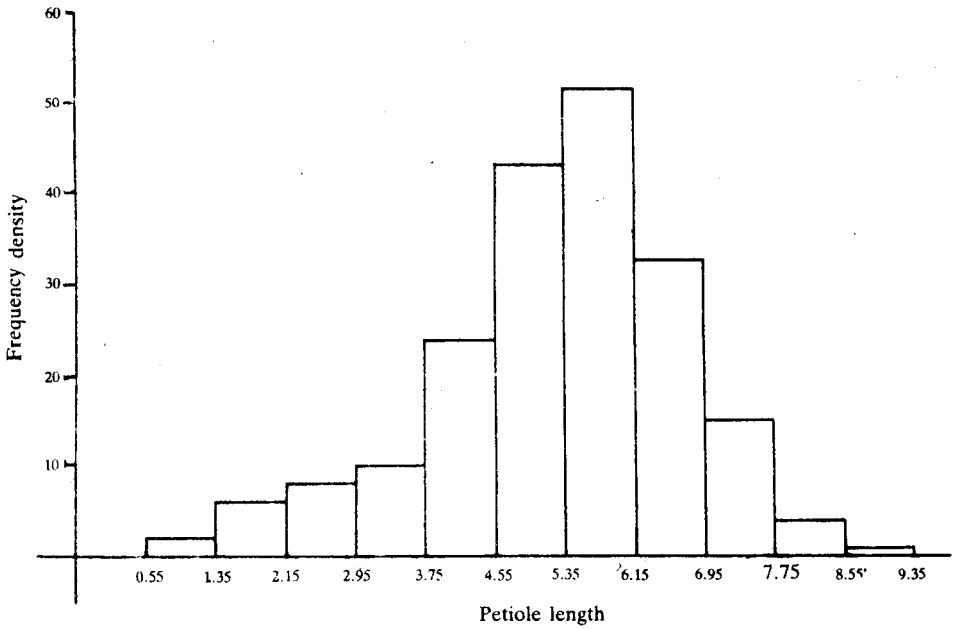


Fig. 7 : Histogram for the frequency distribution of petiole length.

In Fig. 8, you can see the histogram for the frequency distribution of family income given in Table 9.

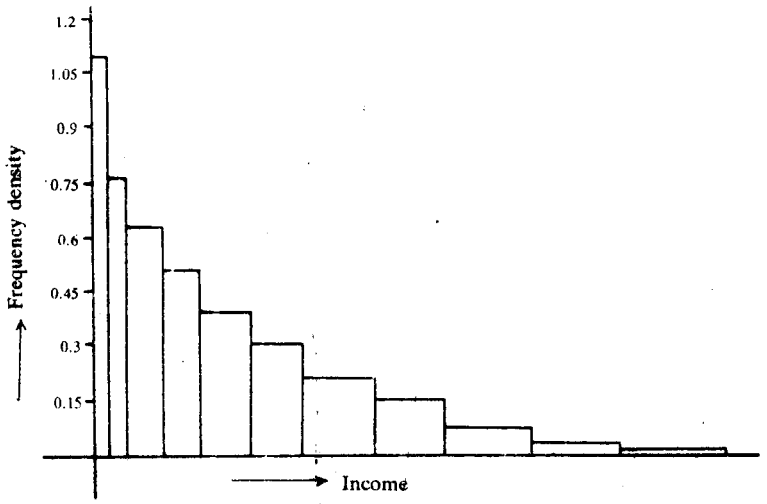


Fig. 8 : Histogram for the frequency distribution of family income

Try this exercise now.

E9) Draw a histogram to represent the frequency distribution of the yield of seed cotton given in E7.

Here we have seen some ways of diagrammatically representing the frequency tables of qualitative characters, and discrete and continuous variables. We can also use cumulative frequencies to represent the frequency distribution of a variable. In the next sub-section, we shall see how the cumulative frequency tables of a variable can be diagrammatically represented.

1.4.2 Cumulative Frequencies

Now we divide our discussion into two parts : a) discrete and b) continuous variables.

a) Discrete variable

We again take two perpendicular axes of coordinates. The vertical axis will now be used for the cumulative frequency while the horizontal axis will continue to be used for the variable itself. But note the way the cumulative frequency changes : in the discrete case, whenever it changes it changes by jumps (a point already mentioned in Sec. 1.3.2).

In Table 4, which is a cumulative frequency table of the less than type, the cumulative frequency is zero for values of the variable less than 1, is 3 for values not less than 1 but less than 2, is 10 for values not less than 2 but less than 3, and so on. Hence, the cumulative frequency diagram takes the form indicated in Fig. 9(a). It is called a **step diagram** owing to its resemblance to a flight of steps.

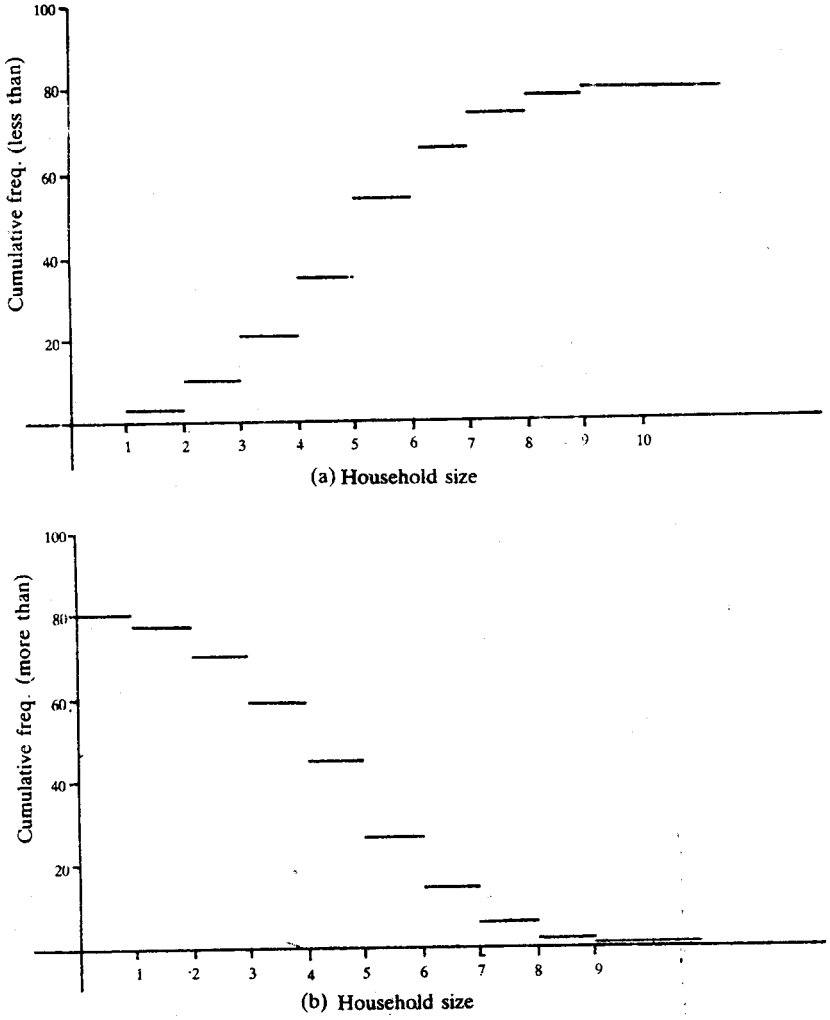


Fig. 9 : Step diagram representing cumulative frequencies of the (a) less than type, (b) more than type, for the data on household size.

The picture takes a somewhat different form when it comes to the cumulative frequencies of the more-than type. From Table 4 you can see that the cumulative frequency diagram will again be a step diagram, but like the one in Fig. 9(b).

b) Continuous variable

In representing the cumulative frequencies of either type for a continuous variable, we proceed as in the discrete case, taking two rectangular axes of coordinates, the horizontal for values of the variable and the vertical for cumulative frequency. But we have to bear in mind that the cumulative frequency of the less than type for any class corresponds to the upper class boundary and that it increases gradually and not by jumps (as it does in the discrete case). Similarly, we have to remember that the cumulative frequency of the more than type for any class corresponds to the lower class boundary and that it decreases gradually and not by jumps.

So while drawing the diagram for the cumulative frequencies of either type, the points corresponding to the successive class boundaries are joined by straight line segments. Note that the cumulative frequency of less than (more than) type is zero (n) for any variable value less than the lower boundary of the lowest class and is n (zero) for any variable value exceeding the upper boundary of the highest class. Hence, the graph of the cumulative frequency of the less than type will coincide with the horizontal axis for values less than the lower boundary of the lowest class and parallel to that axis at a height of n for all values equal to or exceeding the upper boundary of the highest class.

In the case of the cumulative frequency diagram of the more than type, the picture gets reversed : the graph will now be coincident with the horizontal axis for values exceeding the upper boundary of the highest class and will be parallel to that axis at a height of n for all values not exceeding the lower boundary of the lowest class. In Figs. 10(a) and (b), we have these diagrams for the data on petiole length of leaves of a **pipal tree**.

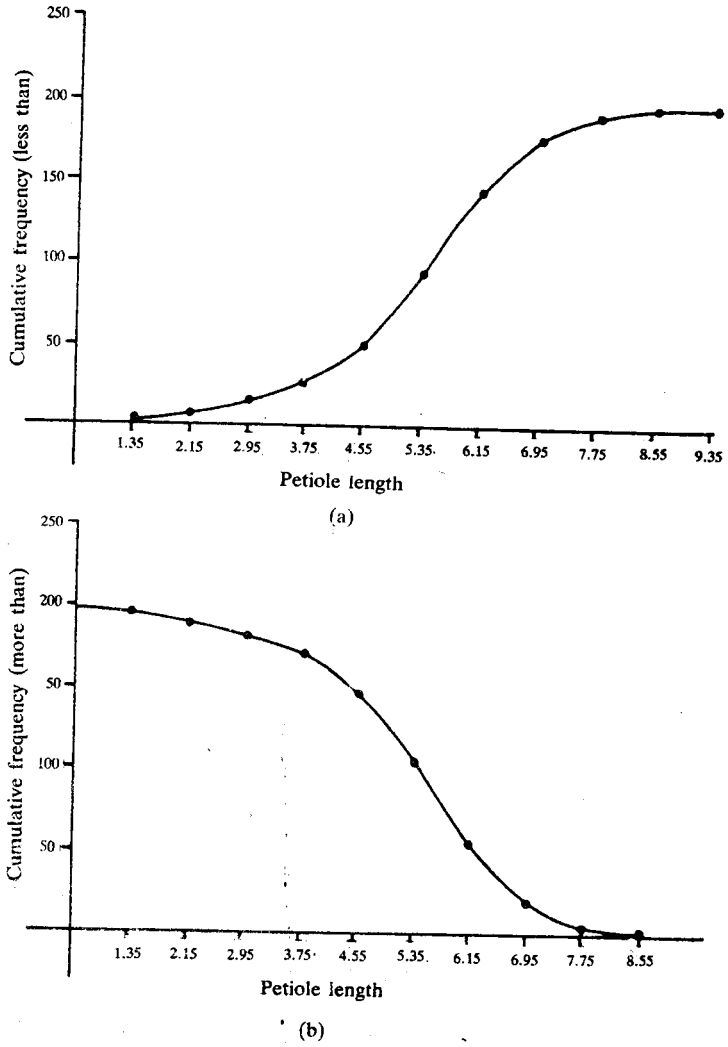


Fig. 10 : Ogives for the data in Table 8 (a) less than type, (b) more than type.

The two cumulative frequency diagrams for a continuous variable resemble in shape the two curves forming the top of an **ogée**, a type of arch. Hence they have been called the **ogives** of the distribution of the variable.

Here is an exercise for you.

E10) a) Represent the cumulative frequencies of the less than and more than types for the data in E6 by suitable diagrams.

b) Draw the ogives corresponding to the data in E7.

So far we have considered frequency distributions of variables where the total number of individuals was finite. Later, in Block 4, you will see that the frequency distributions that we encounter in real life situations arise from sampling from a large group of individuals, called a **population.** In most of these situations, we can regard the population as infinite. Let us now discuss the diagrammatic representation of the frequency distribution of an infinite population by a frequency curve.

1.4.3 Frequency Curve

Let us try to visualise what the frequency distribution or its histogram would look like in an infinite population, especially when the variable is continuous. We first divide [a,b], the range of variation of a continuous variable, into a few class intervals when the total frequency (i.e., the sample size) is small. But let us consider samples of increasing size and at the same time suppose the class intervals are taken smaller and smaller. Suppose we draw the histograms of the distributions obtained in this manner. To make these histograms comparable, we replace frequency density by relative frequency density on the vertical axis. Also see Fig. 11(a), (b) and (c). Isn't it natural then to expect that the histogram will gradually take the form of a smooth curve (Fig. 11 (d))? This smooth curve, representing the frequency distribution of the variable in the infinite population, is called the **frequency curve** of the variable.

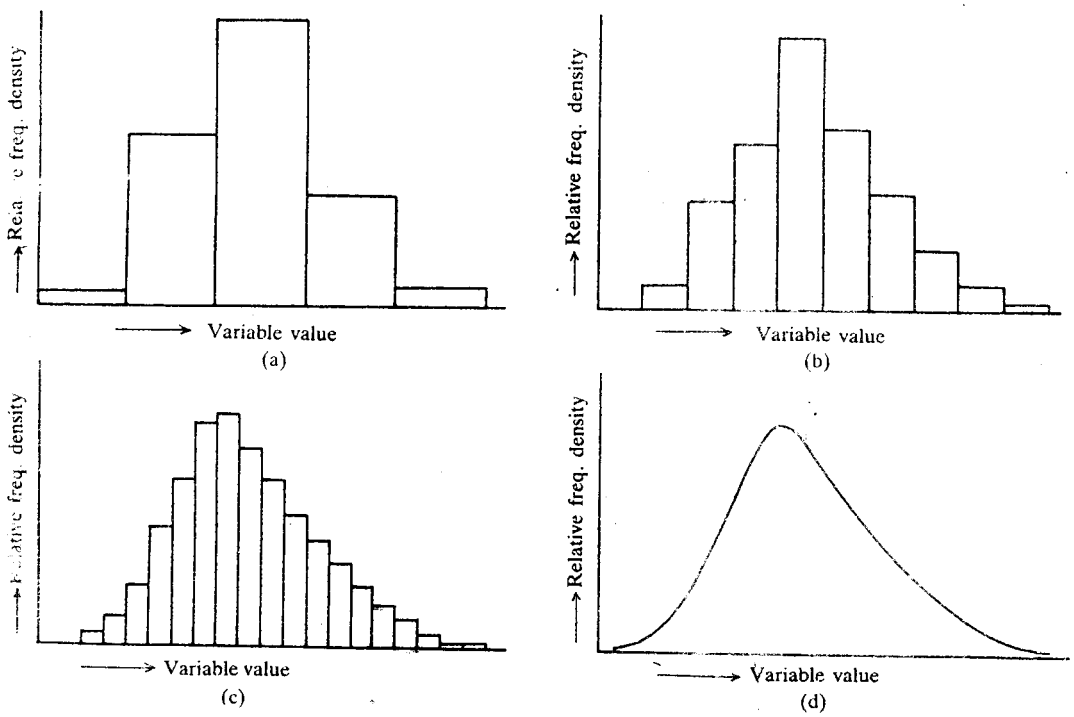


Fig. 11 : Histogram of a frequency distribution of a continuous variable approaching a smooth curve.

Similarly, we can also say that with increasing total frequency and decreasing class width, the ogive of a continuous variable of either type will also gradually approach a smooth curve as shown in Fig. 12. For the sake of comparability, we draw these on the basis of cumulative relative frequencies (rather than cumulative frequencies).

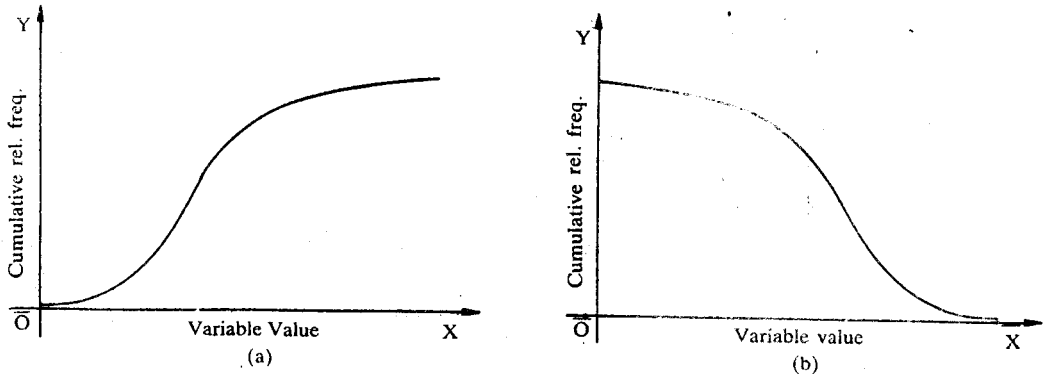


Fig. 12 : Limiting forms of the ogives of a continuous variable : (a) less than type (b) more than type

Now here are some exercises. In each of these we have asked you to give a diagrammatic representation to some of the frequency distributions which you have met in this unit.

E11) Represent the frequency distribution of assessment of the services offered by the nursing home for the 55 inmates (given in E5) in terms of frequencies.

E12) Draw a suitable diagram to represent the frequency distribution given in Table 10.

So far we have discussed various ways of visual representation of frequency distributions. Now we shall see how frequency distributions can be classified into certain broad categories according to shape.

1.4.4 Broad Classes of Distributions

In this section, we'll consider five different classes of distributions. These are

i) Bell-shaped symmetrical
ii) Bell-shaped moderately asymmetrical
iii) J-shaped
iv) U-shaped
v) Multimodal

distributions.

Let's discuss these one by one.

i) Bell-shaped Symmetrical Distribution

Such a distribution is also called a unimodal symmetrical distribution. It may be related to either a discrete or a continuous variable. It has the feature that its highest frequency or frequency density occurs right at the middle of its range of variation, and the frequency or frequency density decreases on either side gradually and at the same rate (see Fig. 13).

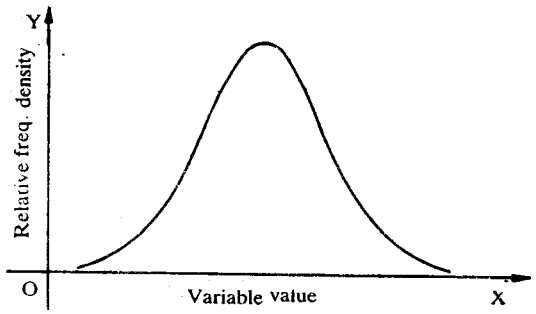


Fig. 13 : Frequency curve of a symmetrical distribution

Many of the distributions that are encountered in the physical, biological and behavioural sciences, as well as those arising from measurements in the field of manufacturing industry, closely follow this form. For instance, if we collect data on the stature (in cm) of a large number of adult males of a given race, then we will end up with a distribution of this type.

ii) Bell-shaped Moderately Asymmetrical Distribution

A distribution of this type also has a single maximum, but the frequency or frequency density decreases on one side at a higher rate than on the other (Fig. 14).

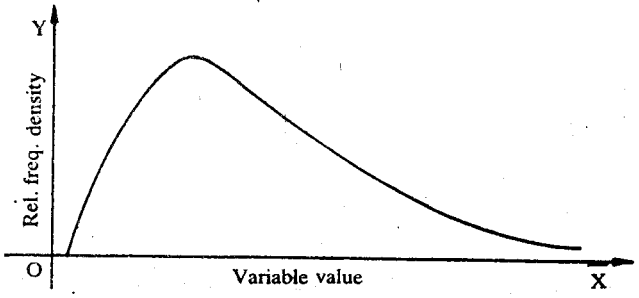


Fig. 14 : Frequency curve of a bell-shaped asymmetrical distribution.

While we very rarely encounter an exactly symmetrical distribution, most of the real-life distributions will fall in the present category. The distribution of petiole length per leaf of a pipal tree, as indicated by the histogram of Fig. 10, has a distribution with a longer tail to the left of the maximum than to the right. The distribution of number of defects per piece of a manufactured item, on the other hand, will have a longer tail to the right of the maximum than to the left. The distribution of the births occurring in a year in a big community by age of mother will also be found to belong to this category.

iii) **J-shaped Distribution**

A J-shaped distribution may be said to be the most extreme form of an asymmetrical distribution. Here the frequency or frequency density is maximum at one end of the range and decreases monotonically as the variable value changes from this end of the range to the other (see Fig. 15).

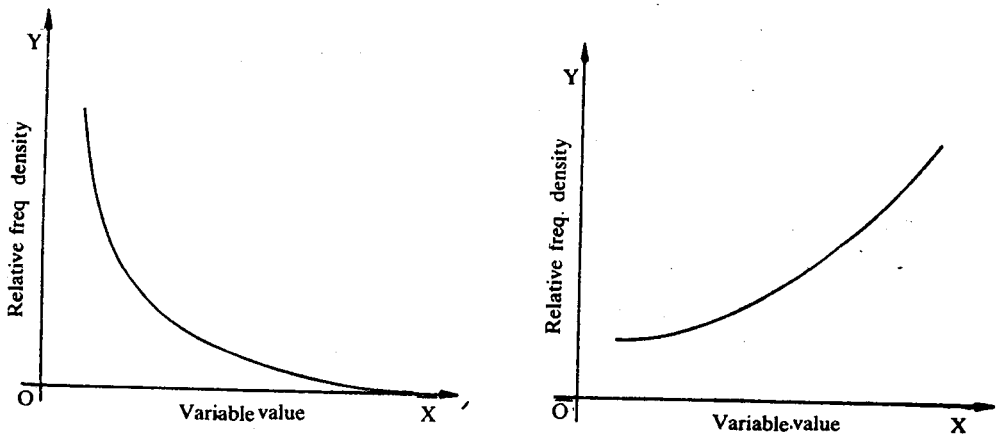


Fig. 15 : Two J-shaped distributions

The income distribution of Table 9 falls in this category, as you can see from Fig. 8.

The distribution of land-holding per family, the distribution of age at death of people of age 60 years or less, the distribution of life of lamp bulbs, etc., will also be similar to Fig. 15 (a).

iv) **U-shaped Distribution**

Such distributions are extremely rare. A distribution of this type has its minimum frequency (or frequency density) towards the middle of the range of variation while the frequency (or frequency density) gradually increases, at the same rate or at different rates, as the variable value changes either to the left or to the right (see Fig. 16).

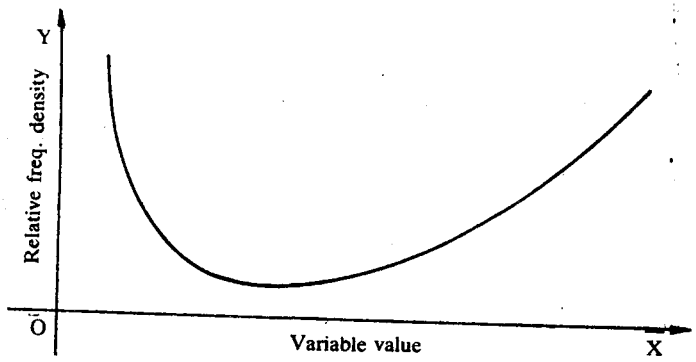


Fig. 16 : A U-shaped distribution.

The distribution of days in a month by the degree of cloudiness at a place (if cloudiness may be considered a continuous variable) has been found to follow this form. In other words, the number of days with no or very high cloudiness will be large, while there will be fewer days with moderately low or moderately high degrees of cloudiness.

v) **Multimodal Distribution**

In some situations, we may come across distributions with more than one maximum as in Fig. 17. You may realise that such a distribution may result if several groups of individuals are mixed together. For each group separately, the distribution may be unimodal, but if they have distinct maxima, then the distribution in the composite group will take on a multimodal form.

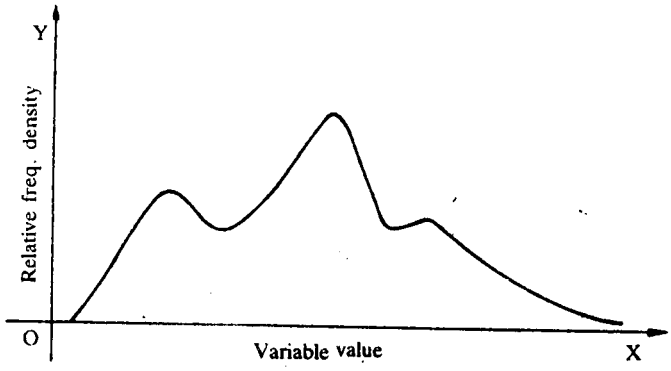


Fig. 17 : A multimodal distribution

In general, a multimodal distribution signifies heterogenity of the data—the fact that the data have been obtained from groups with widely different characteristics.

So, we have seen that we can obtain a lot of information about the data from its pictorial representation.

With this we bring this unit to a close. Let us go back and recall the points covered in it.

1.5 SUMMARY

In this unit, we have discussed the following points :

1) The term **statistics** may mean either numerical data arising in some sphere of life or the scientific discipline that concerns itself with the collection, analysis and interpretation of such numerical data.
2) Methods of data collection :
   Direct observation method
   Questionnaire method
   Interview method
3) Classification of characters into **qualitative** and **quantitative**, and that of quantitative characters into **discrete** and **continuous** ones.
4) Representation of frequency distribution of a character by means of a table.
5) Relative frequencies and cumulative frequencies.
6) Representation of the frequency distribution of a character by means of a diagram : bar diagram, pie diagram, column diagram, frequency polygon, histogram, ogive curve.
7) Classification of univariate distributions into certain broad categories :
   Bell-shaped symmetrical and asymmetrical distributions,
   J-shaped distributions,
   U-shaped distributions,
   Multimodal distributions.

1.6 SOLUTIONS

E1) a) and d) are secondary data,
   b) and c) are primary.

E2) a) interview method
b) questionnaire method (or interview method)
c) measurement (of yield for some sample plots)
d) measurement.

E3) c) and d) are qualitative,
a), b) and e) are quantitative.

E4) b), c) and f) are discrete,
a), d) and e) are continuous.

E5) The frequency distribution is :

Assessment	Frequency	Relative Frequency
V	7	0.1272
G	16	0.2909
S	18	0.3272
B	13	0.2363
P	1	0.0181
	55	

a) 16

b) 41

c) $\frac{14}{55} \times 100 = 25.4545\%$

E6)

Word Length	Frequency	Relative Freq.	Cumulative frequency less than	Cumulative frequency more than
2	13	0.1428	13	91
3	19	0.2087	32	78
4	21	0.2307	53	59
5	15	0.1648	68	38
6	9	0.0989	77	23
7	6	0.0659	83	14
8	4	0.0439	87	8
9	3	0.0329	90	4
10	1	0.0109	91	1
	91			

a) 77

b) 38

c) 0.1428

d) $\frac{59}{91} = 0.6483$

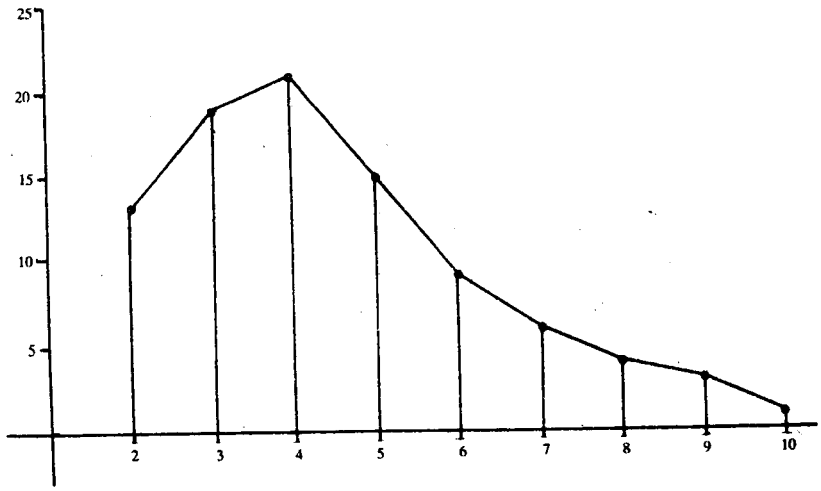
E7) Noting that the lowest and the highest of the observations are 23 and 115, you may take your classes as 21-30, 31-40,....., 111-120.
(a)

Yield (gm)	Class mark	Frequency	Relative Freq.	Cumulative frequency less than	Cumulative frequency more than
21—30	25.5	2	0.0166	2	120
31—40	35.5	5	0.0416	7	118
41—50	45.5	13	0.1083	20	113
51—60	55.5	21	0.175	41	100
61—70	65.5	27	0.225	68	79
71—80	75.5	22	0.1833	90	52
81—90	85.5	16	0.1333	106	30
91—100	95.5	9	0.075	115	14
101—110	105.5	4	0.0333	119	5
111—120	115.5	1	0.0083	120	1
		120			

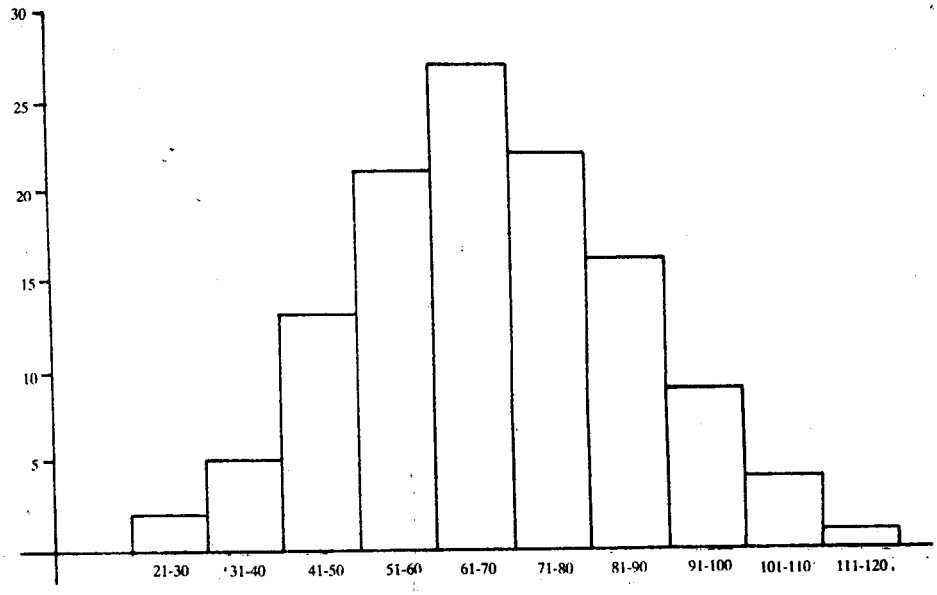
b) i) 43.5 ii) 5 iii) 41

c) i) $\frac{47}{120}$ ii) $\frac{79}{120}$ iii) $\frac{3}{120}$

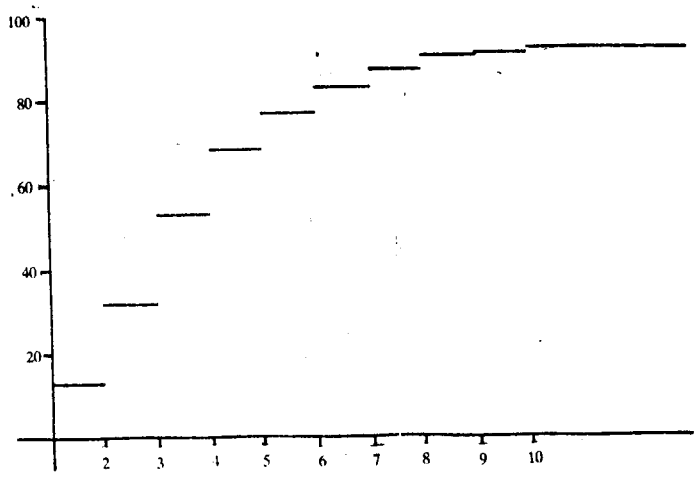
E8)

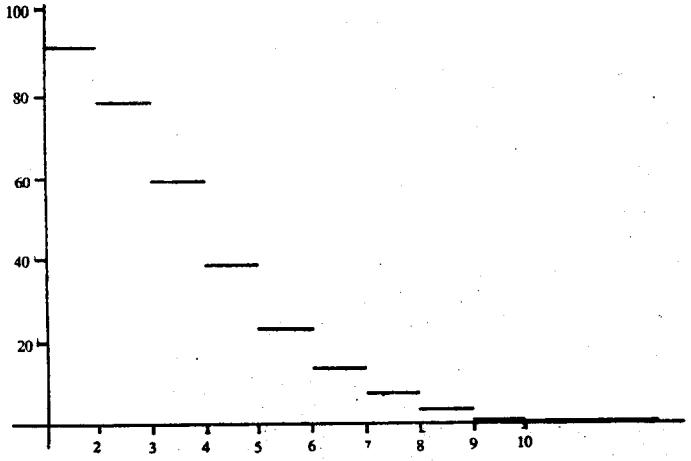


E9)

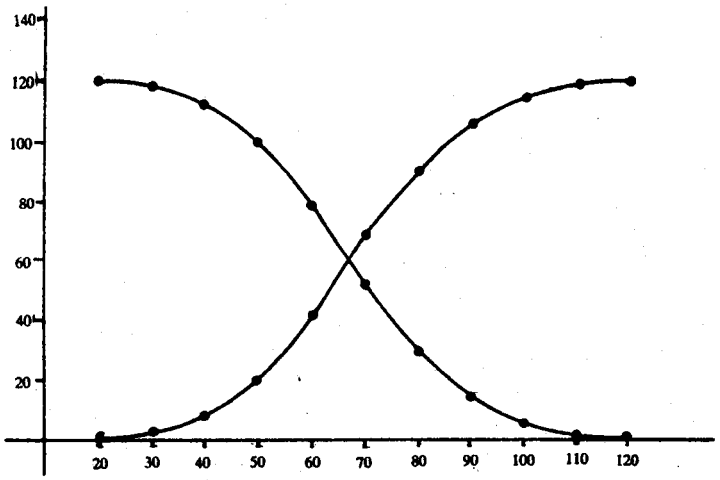


E10) a)

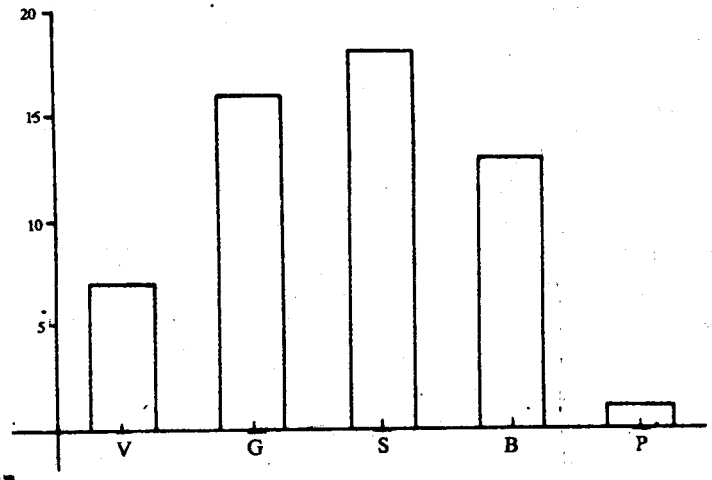



100
80
60
40
20
2
3
4
5
6
7
8
9
10

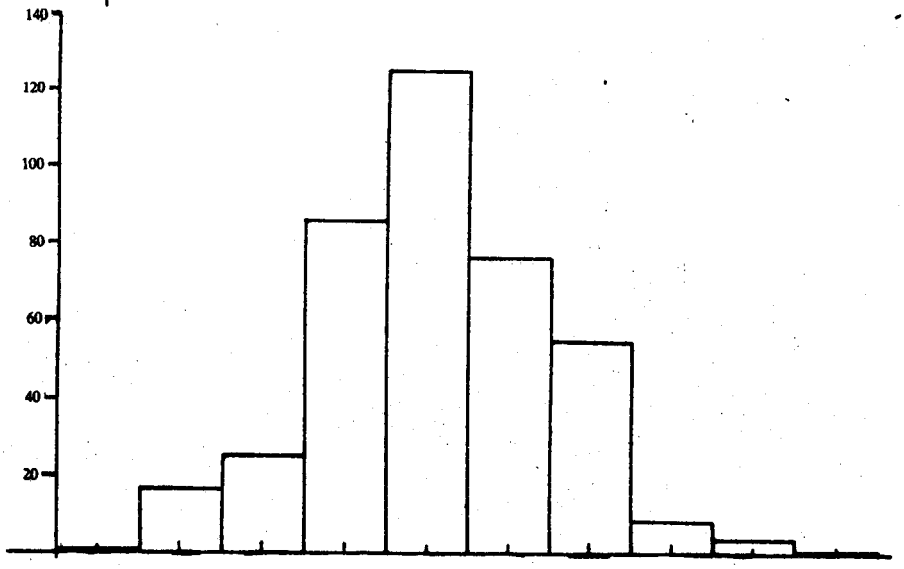
E10) b)


140
120
100
80
60
40
20
20
30
40
50
60
70
80
90
100
110
120

E11)


20
15
10
5
V
G
S
B
P

E12)


140
120
100
80
60
40
20

UNIT 2 MEASURES OF CENTRAL TENDENCY AND DISPERSION

Structure

2.1 INTRODUCTION

In Unit 1, we have seen that statistical data may relate to qualitative characters as well as quantitative. This unit highlights some common features of a frequency distribution of a single variable (also called a univariate frequency distribution). These features are central tendency and dispersion. We shall also discuss some commonly used measures of these features and their properties. Most of you would already be familiar with the measures of central tendency and dispersion. So we'll go over these quickly and ask you to do a few exercises to help you to recapitulate.

In the examples, we'll be referring again and again to the data sets which you have studied in Unit 1. So, you will often have to go back and look at the tables in that unit. If you have a calculator, it would be a good idea to keep it handy while going through this unit. Calculators are also available at your study centre.

Now we are going to list the objectives of this unit. After you have gone through the unit, make sure that you have achieved them.

Objectives

After studying this unit, you should be able to :

- compute the mean, median and mode from raw data or from a given frequency distribution,
- compute the range, standard deviation and mean deviation of the data, whether grouped or ungrouped.
- derive and use some algebraic properties of the measures of central tendency and dispersion.

2.2 CENTRAL TENDENCY AND DISPERSION

If you glance through any set of observations on a variable, you will usually find a tendency among the observations to cluster around some particular point, or some small part, of the range of variation. This tendency will be all the more apparent if we construct the column diagram (or frequency polygon) or histogram related to the data. In Fig. 1, you can see the observations clustering in the interval [4.55—6.15].

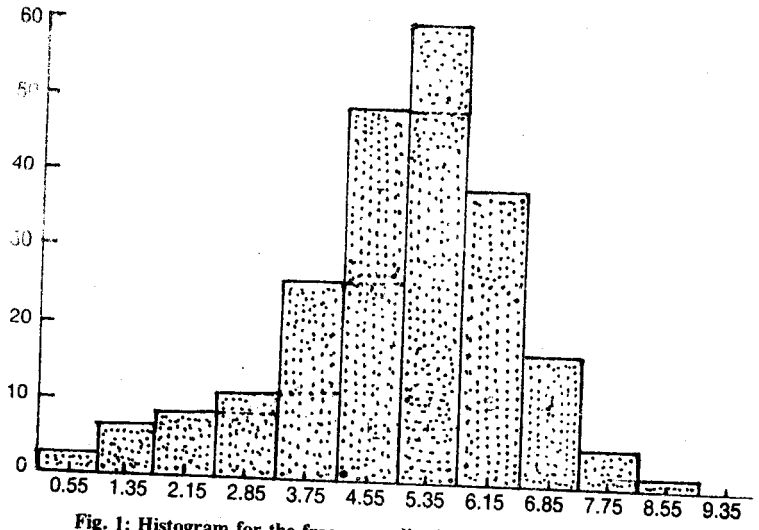


Fig. 1: Histogram for the frequency distribution of petiole length

This tendency is called the **central tendency** of the variable (or of its frequency distribution). In the presence of central tendency, we can take the particular value, or a value in the particular part of the range, around which the observations cluster, as representative or typical of the whole set. By a **measure of central tendency** (or a measure of location) of a frequency distribution, we mean such a typical value. The same is meant by the more familiar term, **average.** Note that any such measure must have a unit. This unit corresponds to the unit in which the variable is measured and recorded.

Now suppose you are told that the average blood pressure for your age-group is 120 mm. You measure yours and find that it is 110 mm. What does this mean? Is it a cause for worry? Or does your blood pressure fall within the limits of normal **variation**? So, you see, knowing only the 'average' is not enough. While mentioning an average, we should also give an idea of the extent to which the individual observations differ from the average value. The variation of the observations from the average (or from one another) is called **scatter** or **dispersion** of the data on the variable. Thus, to describe a set of data, we have to give a measure of dispersion along with a measure of central tendency.

In the next section, we'll describe some measures of central tendency. But before that, a word about the notation that we'll be using in the unit.

Notation

We shall use some such letter as x, y or z to denote the variable under study, and the letter n to denote the sum total of the frequencies (i.e., the total number of individuals for which data are available).

In case the data are in their raw (or ungrouped) form, x_i will denote the values of x as observed for the ith individual (i=1, 2, ... n). Thus, x_1, x_2, x_3, etc., will, respectively, denote the values of x as observed in the first individual, second individual, third individual, etc.

In case the data are in the form of a frequency table, the number of classes will be denoted by k and x_i will denote the value of x defining the ith class or the mid-point of the ith class interval (also called the ith **class mark**). We will denote the frequency in the ith class by f_i. We then have

$$\sum_{1}^{k} f_i = n. \qquad \ldots (1)$$

Before we go any further, let us recall the following properties of the summation notation.

i) If $a_i=a$ for all i, then

$$\sum_{i=1}^{m} a_i = ma.$$

ii) If b is a constant, then

$$\sum_{i=1}^{m} bx_i = b\sum_{i=1}^{m} x_i$$

iii) $\sum_{i=1}^{m}(x_i+y_i) = \sum_{i=1}^{m} x_i + \sum_{i=1}^{m} y_i$, $\sum_{i=1}^{m}(x_i-y_i) = \sum_{i=1}^{m} x_i - \sum_{i=1}^{m} y_i$.

This result can be extended to more than two series of numbers.
i), ii) and iii) lead to the following property:

iv) If a and b are constants, then

$$\sum_{i=1}^{m}(a+bx_i) = ma + b\sum_{i=1}^{m} x_i.$$

In case the observations $x_1, x_2, \ldots\ldots, x_n$ on a variable x are arranged in ascending order, we shall denote by $x_{(i)}$, the ith value in this arrangement.

Thus, we have

$$x_{(1)} \leq x_{(2)} \leq \ldots. \leq x_{(n)} \qquad \ldots (2)$$

In the same way, if the observations are arranged in descending order, then $x'_{(i)}$ may be taken to denote the ith value in that arrangement. Thus, we have

$$x'_{(1)} \geq x'_{(2)} \geq \ldots. \geq x'_{(n)} \qquad \ldots (3)$$

Our notation also implies that f_i/n will stand for the relative frequency of the ith class (i=1, 2,...,k). We denote by F_i the cumulative frequency of the less than kind and by F'_i the cumulative frequency of the more than type for the ith class. By definition, then,

$$F_i = \sum_{j=1}^{i} f_j \quad (i=1, 2,\ldots..,k) \text{ and}$$

$$F'_i = \sum_{j=i}^{k} f_j \quad (i=1, 2,\ldots..,k),$$

so that

$$F_i = F_{i-1} + f_i \quad \text{for } i \geq 2,$$

and

$$F'_i = F'_{i+1} + f_i \quad \text{for } i \leq k-1$$

Now here is a simple exercise to find out whether you have understood our notation or not.

E1) a) Suppose the annual incomes of 5 individuals as reported in the I-T returns for the year 1990-91 (in thousands of rupees) are 75, 80, 75, 105 and 83. Compute $x_{(3)}$ and $x'_{(2)}$.

b) Show that

$$F'_i = n - F'_{i+1} \quad (\text{for } i \leq k-1)$$

and

$$F'_i = n - F_{i-1} \quad (\text{for } i \geq 2)$$

c) What are F_k and F'_1 ?

Now we are ready to discuss the measures of central tendency in the next section.

2.3 MEASURES OF CENTRAL TENDENCY

The averages (or measures of central tendency) in common use are mean, median and mode. Let's consider these one by one.

2.3.1 The Mean

The arithmetic mean (or simply the mean) is the most widely used average. For any set of data on the variable x, the mean is denoted by $\bar{x}$ and is obtained by dividing the sum of the observations by their number. Thus, we have

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i \qquad \ldots (4)$$

For data grouped into a frequency table, we have

$$\bar{x} = \frac{1}{n}\sum_{1}^{k} f_i x_i \qquad \ldots (5)$$

Formula 5 will provide the exact answer in case the variable x is discrete and the frequency table has classes defined by one distinct value of x each.

If classes of the frequency table are not defined by single distinct values of x, we still use Formula 5, but x_i now denotes the **class mark** of the ith class (i=1, 2,....., k). Since we are considering class marks instead of individual values, Formula 5 gives only an approximate value of $\bar{x}$. In other words, we can say that Formula 5 is subject to grouping errors. But Formula 4 gives the value free from such errors. In the continuous case, since the observations will involve rounding-off errors, even Formula 4 will be approximate, despite being free of grouping errors.

We now give two examples to illustrate the use of Formulas 4 and 5.

Example 1 : The quantities of milk (in litres) produced by a dairy farm on ten consecutive days are shown below :

218.2	199.7	207.3	185.4	213.7
184.7	179.5	194.4	224.3	203.5

Let us calculate the mean production.

Here the data concern a continuous variable, viz., milk yield per day. Here n = 10 and

$$\sum_{1}^{n} x_i = 2010.7 \text{ litres.}$$

Hence, the mean output per day for the dairy farm is

$$\bar{x} = \frac{2010.7}{10}$$
$$= 201.07 \text{ litres.}$$

Before giving the next example, we would like to tell you about a method of simplifying the computation of the mean.

In Sec. 2.3.4, you will see that under a linear transformation of the variable, the mean gets transformed in the same way. Hence, if the variable is subjected to a change of base and/or scale, that is, if

$$u = \frac{(x-x_0)}{c} \ (c \neq 0), \quad \text{then} \quad \bar{u} = \frac{(\bar{x}-x_0)}{c}$$

Thus, we get $\bar{x} = x_0 + c\bar{u}$ $\qquad \ldots (6)$

We can use this formula to simplify the computation of the mean from a frequency table. We take x_0 to be the class mark of a class near the middle of the table, and c to be the common width of the classes. Example 2 may help you in getting this point clear.

Example 2 : Let us obtain the mean petiole length per leaf of pipal tree from Table 6 in Unit 1.

We lay out the computations as in the table below :

Table 1 : Calculations for mean petiole length

Class mark x_i	$u_i=(x_i-4.95)/0.8$	Frequency f_i	f_iu_i
0.95	−5	2	−10
1.75	−4	6	−24
2.55	−3	8	−24
3.35	−2	10	−20
4.15	−1	24	−24
4.95	0	43	0
5.75	1	52	52
6.55	2	33	66
7.35	3	15	45
8.15	4	4	16
8.95	5	1	5
Total	—	198	82

From the table, we have, as the mean of the new variable u,

$$\bar{u} = \frac{1}{n}\sum_{1}^{k} f_i u_i$$

$$= \frac{82}{198} = 0.4141.$$

Since $u = \frac{(x-4.95)}{0.8}$, the mean petiole length is

$$\bar{x} = 4.95+0.8\,\bar{u}$$

$$= 4.95+0.331=5.281 \text{ cm.}$$

You can try your hand at these exercises now.

E2) The scores obtained in English by 15 students are given below. Calculate the mean score.

33, 41, 46, 47, 52, 52, 53, 54, 57, 61, 61, 68, 69, 70, 74

E3) The age-distribution of the Indian population according to the 1981 census is shown below :

Age as on last birthday	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-74
percentage	12.59	14.08	12.88	9.63	8.62	7.63	6.38	5.85	5.14	4.40	3.83	2.47	6.49

Obtain the mean age of an Indian alive at the time of the census (Note that here age x on last birthday means that the age at the time of the census was less than x+1 years but not less than x years. Hence the class intervals should be taken to be 0-5, 5-10, etc. You can take $x_0=32.5$ and c=5.

Actually, the census tabulation leaves the last interval open, that is, simply as "60−" But here we ask you to do the calculation based on the assumption that the upper end point of this interval is 74. We cannot calculate the mean without some such assumption. However, we must realise that this assumption may introduce yet another source of error.

We now turn our attention to another measure of location, the median.

2.3.2 The Median

Refer to Sec. 2.2 for notation.

By the median of a set of observations on a variable x, we mean the middlemost value of the set when the elements are arranged in either ascending or descending order. So there are at most half the observations below the median as well as above the median.

But the middlemost value may not be unique, and so the median too may not be unique. Let us denote the median by $\tilde{x}$. For ungrouped data arranged in ascending order, we define $\tilde{x}$ as follows.

If n is odd (=2m+1, say), $\tilde{x} = x_{(m+1)}$.

If n is even (=2m), any value $\tilde{x}$ such that $x_{(m)} \leq \tilde{x} \leq x_{(m+1)}$ is a median of x.

Of course, if $x_{(m)}$ and $x_{(m+1)}$ are equal, then median will be unique even when n is even. If n is even and $x_{(m)} \neq x_{(m+1)}$, we sometimes take the mean of these two central values as the unique median of x, i.e. (by convention), we take

$$\tilde{x} = \frac{x_{(m)} + x_{(m+1)}}{2}.$$

In case the data are arranged in descending order, we have

$\tilde{x} = x'_{(m+1)}$ if n is odd (=2m+1)

and $x'_{(m+1)} \leq \tilde{x} \leq x'_{(m)}$ if n is even (=2m)

Here too, if n=2m and $x'_{(m)} \neq x'_{(m+1)}$, we sometimes take

$$\tilde{x} = \frac{x'_{(m)} + x'_{(m+1)}}{2}.$$

Example 3 : Consider the data on the daily milk yield of a dairy farm that were cited in Example 1. Arranged in ascending order, the observations are (in litres),

179.5 184.7 185.4 194.4 199.7 203.5 207.3 213.7 218.2 224.3

Here n is even (=10). Also, $x_{(5)}$=199.7 and $x_{(6)}$=203.5.

Hence, any value between 199.7 litres and 203.5 litres may be taken to be the median yield of milk for the dairy farm. However, if we follow the convention, we may take, as the unique median,

$$\tilde{x} = \frac{(199.7+203.5)}{2} = 201.6 \text{ litres}$$

Now suppose the data on a discrete variable is put in the form of a frequency distribution in which each class is defined by a single value of x. In this case, the cumulative frequency table of less than (more than) type presents an arrangement of the original observations in ascending (descending) order. Let's see how we can use this fact to get the median.

Example 4 : Table 4 a in Unit 1 shows that if the original data (as shown in Table 2 of Unit 1) were arranged in ascending order, then the first three values would be 1, the 4th to the 10th would be 2, the 11th to the 21st would be 3, and so on. Here n=80 and we find that

$x_{(40)} = x_{(41)} = 5$,

so that $\tilde{x} = 5$.

When the variable x is continuous and the data are in grouped form, you may visualise the frequency curve of the distribution. The median should then be taken as the value of x that divides the area under the frequency curve into two equal parts (or the value that has ordinate $\frac{n}{2}$ in the corresponding ogive of either type). In any particular situation, however, you will have a frequency table with, say, at most 20-25 class intervals, and can hope to get only a rough approximation to the median. You may then take the median as that value of x which has cumulative frequency (of either type), n/2.

Suppose we have a cumulative frequency table of less than type. We first ascertain which of the class-intervals contains the median. Suppose this interval has lower boundary x_l and upper boundary x_u. We further assume that the cumulative frequency increases linearly from F_l to F_u as the variable x increases from x_l to x_u.

Thus, $\frac{\tilde{x}-x_l}{x_u-x_l} = \frac{(n/2)-F_l}{F_u-F_l}$,

so that $\tilde{x} = x_l + \frac{(n/2)-F_l}{f_0} \times c$, ... (7)

where $c=x_u-x_l$ is the width of the median interval and $f_0 = F_u-F_l$ is the frequency for the interval.

We can also make use of the cumulative frequency table of the more than type for computing the median.

Here is an example to illustrate the use of Formula (7).

Example 5 : For the frequency distribution of petiole length per leaf of a pipal tree, $\frac{n}{2} = 99$. Therefore, Table 8 in Unit 1 indicates that the median would be in the interval 5.35 cm – 6.15 cm. Thus,

$$x_l = 5.35,\ x_u = 6.15 \Rightarrow c = 0.8$$

and

$$F_l = 93,\ F_u = 145 \Rightarrow f_0 = 52.$$

We have assumed that the ratio in which $\tilde{x}$ divides $[x_l, x_u]$ is the same as the ratio in which n/2 divides $[F_l, F_u]$.

Hence, the median may be taken to be

$$\begin{aligned}\tilde{x} &= 5.35 + \frac{99-93}{52} \times 0.8 \\ &= 5.35 + 0.009 \\ &= 5.359 \text{ cm.}\end{aligned}$$

We can also approximate $\tilde{x}$ graphically.

Let us see how.

After drawing the ogive of either type, draw a line parallel to the horizontal axis (i.e., x-axis) at a height, n/2. The abscissa of the point at which the line intersects the ogive is $\tilde{x}$. Note that $\tilde{x}$ is also the abscissa of the point of intersection of the two ogives, see Fig. 2(a) and (b).

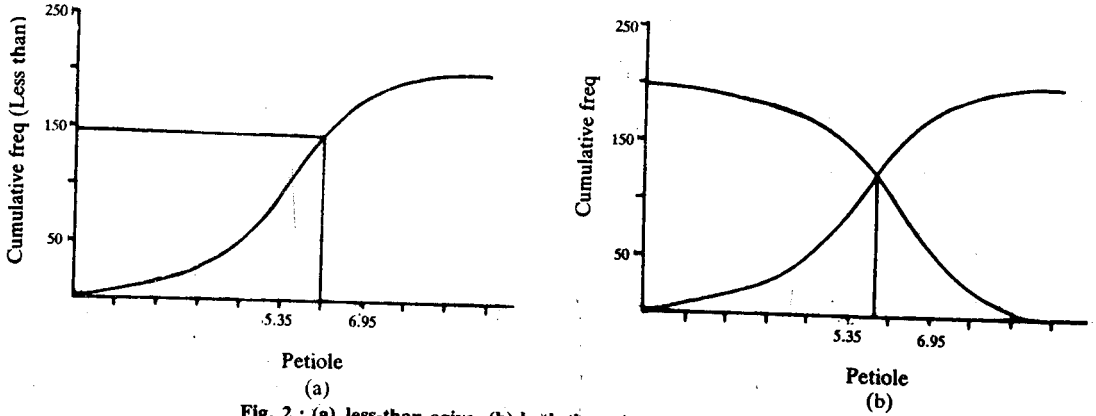


Fig. 2 : (a) less-than ogive, (b) both the ogives for the data on petiole length.

You may use the same method in determining the median for the frequency distribution of a discrete variable like the one in Table 10 in Unit 1, taking the artificial class intervals 7.5-12.5, 12.5-17.5, etc. But note that this method can provide us with only a rough estimate of the median. This is because in this method we only replace a step diagram by an ogive.

We have described the methods used to compute the median of

i) raw data on a discrete variable,

ii) data in the form of an ungrouped frequency table, and

iii) data in the form of a grouped frequency table.

On the basis of our discussion, see if you can solve the following exercise.

E4) Obtain the median age of an Indian according to the population census of 1981 on the basis of the frequency table given in E3.

Now we take up the third common measure of central tendency—the mode.

2.3.3 The Mode

In French **la mode** means the fashion. By the mode of a set of observations, we also mean the fashionable value, or the value that occurs most frequently, in the set. We shall denote it by $\overset{\circ}{x}$. Mode is an especially useful measure of central tendency for data on purely qualitative characters such as preferences for colours. This is because, in such cases, no other measure would be meaningful.

If the variable is discrete, you may just draw up a frequency table corresponding to each individual value and see which, if any, of the values defining the classes has the highest frequency. If there is such a value, then it will be the mode $\overset{\circ}{x}$. For example, consider the data on household size for 80 households given in Table 3 of Unit 1. The value 5 has the highest frequency. Hence, here the mode is

$$\overset{\circ}{x}=5.$$

However, you may encounter cases where two or more distinct values of the variable have the same frequency, which is higher than the frequencies for the other values. In such cases, we shall say that the mode is not uniquely defined.

Here is an example of a situation where the mode is not uniquely defined.

Example 6 : For the frequency distribution of word length for the 91 words in a poem, as shown in Table 2, you can see that both 4 and 5 have the highest frequency (19). As such, in this case the mode is not unique.

Table 2 : Frequency table of word length for the 91 words in a poem

Wordlength	Frequency
2	13
3	17
4	19
5	19
6	9
7	6
8	4
9	3
10	1
Total	91

When it comes to a grouped frequency distribution, the above approach is inappropriate. This is because now the frequencies correspond to certain class-intervals rather than to single values of the variable. If the classes are of equal width, we may talk of the **modal class** as being the class with the highest frequency (if such a class exists). But since the classes may be of varying width, it is more appropriate to say that the modal class, if it exists, is the class with the highest frequency density (or relative frequency density).

In the case of a continuous variable, we consider the frequency curve which can be obtained as a limiting form of the histogram by taking finer and finer classes and increasing the total frequency at the same time. Then the mode is the value (if there is any), with the highest relative frequency density in the frequency curve (see Fig. 3).

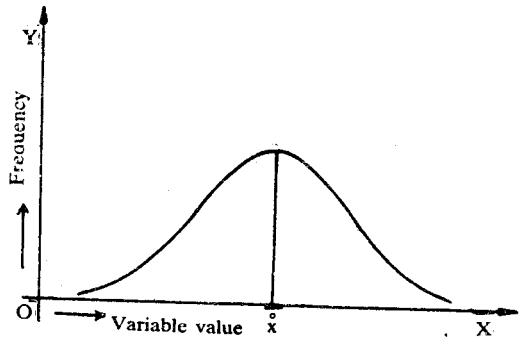


Fig. 3 : A frequency curve with its mode, $\overset{\circ}{x}$.

So, to find the mode of a continuous variable, we look at its frequency table or histogram. As a first approximation, we can take the mid-point of the modal class as $\overset{\circ}{x}$. If this class has boundaries x_l and x_u, and $\overset{\circ}{x}^{(1)}$ denotes the first approximation to $\overset{\circ}{x}$, then

$$\overset{\circ}{x}^{(1)} = \frac{(x_l + x_u)}{2} \qquad \ldots (9)$$
$$= x_l + c/2,$$

where c is the width of the modal class.

But we get a better approximation if we consider the modal class as well as its two adjacent classes (provided, of course, that the modal class is not a terminal class). Suppose that these three classes are of the same width. Let us denote by f_0, f_- and f_+, respectively, the frequency of the modal class, that of the class immediately preceding the modal class and that of the class following the modal class. Further, we assume that

$$\overset{\circ}{x} - x_l : x_u - \overset{\circ}{x} = f_0 - f_- : f_0 - f_+$$

This assumption leads to the second approximation, viz.,

$$\overset{\circ}{x}^{(2)} = x_l + \frac{f_0 - f_-}{(f_0 - f_-) + (f_0 - f_+)} \times c$$
$$= x_l + \frac{f_0 - f_-}{2f_0 - f_- - f_+} \times c. \qquad \ldots (10)$$

Now, when will these two approximations be equal? On equating (9) and (10) and simplifying, we get

$\overset{\circ}{x}^{(2)}$ is equal to $\overset{\circ}{x}^{(1)}$ iff $f_- = f_+$.

In the following example, we have calculated $\overset{\circ}{x}^{(1)}$ and $\overset{\circ}{x}^{(2)}$ for the data in Table 7 in Unit 1.

Example 7 : The frequency table for petiole length per leaf for 198 leaves of a pipal tree has classes of equal width and the class 5.35–6.15 (cm) is the modal class.

Our first approximation to the mode is, then,

$$\overset{\circ}{x}^{(1)} = 5.35 + 0.8/2$$
$$= 5.75 \text{ cm.}$$

In this case, we have

$f_0 = 52$, $f_- = 43$ and $f_+ = 33$,

Thus, the second approximation is

$$\overset{\circ}{x}^{(2)} = 5.35 + \frac{52 - 43}{(2 \times 52) - 43 - 33} \times 0.8$$
$$= 5.35 + \frac{9}{28} \times 0.8$$
$$= 5.35 + 0.257$$
$$= 5.607 \text{ cm.}$$

Empirical = based on practical experience.

Now, here is a remark about the relationship between the three measures that we have discussed.

Remark 1 : There is an **empirical** relation connecting the mean, median and mode, of a distribution, viz., the relation

$$\text{mean} - \text{mode} = 3(\text{mean} - \text{median})$$

or
$$\bar{x} - \overset{\circ}{x} = 3(\bar{x} - \tilde{x}) \qquad \ldots (11)$$

We can also use this in obtaining an approximate value of the mode of a frequency distribution. From (11), we get the formula for this third approximation as,

$$\overset{\circ}{x}^{(3)} = 3\tilde{x} - 2\bar{x}.$$

We are now giving an exercise which will give you some practice in calculating the mode.

E5) a) Find approximately the mode of the age distribution given in E3 by using Formula (10).

b) Compare the mean, median and mode as obtained by you for this distribution and state whether the empirical relation (11) is borne out by the distribution.

So far we have acquainted you with three measures of central tendency; the mean, the median and the mode. In the next sub-section, we shall discuss some algebraic properties of these measures.

2.3.4 Algebraic Properties of the Measures

The mean, median and mode have certain algebraic properties. We are going to list and prove some of them here. You should keep these in mind while using any of these measures.

i) If all the observations on the variable x are equal, say to a, then

$$\bar{x} = \tilde{x} = \mathring{x} = a.$$

ii) If $y = a+bx$ $(b \neq 0)$, then the mean, median and mode of y are

$$\bar{y} = a + b\bar{x},\ \tilde{y} = a + b\tilde{x},\ \mathring{y} = a + b\mathring{x}.$$

iii) The sum of the deviations of the observations on a variable x from their mean is zero i.e.,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

The difference $x_i - a$ is called the **deviation** of x_i from a.

iv) Suppose k sets of observations on x are combined, the ith set having n_i observations with mean $\bar{x}_i$. Then the composite (or grand) mean of x is

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$

v) If $z_i = x_i + y_i$, then $\bar{z} = \bar{x} + \bar{y}$.

Similarly, if $z_i = x_i - y_i$, then $\bar{z} = \bar{x} - \bar{y}$.

This result can be extended to the case when two or more variables are added or subtracted. So, if

$$z = a+bx+cy+dz+\ldots+lw, \text{ then } \bar{z} = a+b\bar{x}+c\bar{y}+d\bar{z}+\ldots..+l\bar{w}.$$

Out of these, i) and v) are easy to prove. We are sure you will be able to prove them. Here, we will prove ii), iii) and iv).

Proof : ii) We have

$$y_i = a + bx_i \text{ for each } i,\ i=1,2,\ldots,n.$$

$$\Rightarrow \sum_{1}^{n} y_i = \sum_{1}^{n}(a + bx_i)$$

$$= na + b\sum_{1}^{n} x_i$$

$$\Rightarrow \bar{y} = a + b\bar{x}.$$

Again, we shall have, for each i,

$$y_{(i)} = a + bx_{(i)} \text{ if } b > 0$$
$$= a + bx'_{(i)} \text{ if } b < 0.$$

In other words, the smallest observation on y corresponds to the smallest observation on x, second smallest observation on y corresponds to the second smallest on x, ... On the other hand, in case $b < 0$, the smallest observation on y

corresponds to the largest observation on x. The second smallest observation on y corresponds to the second largest on x, and so on.

Now if $n = 2m+1$, then $y_{(m+1)} = a+bx_{(m+1)}$ for all $b \neq 0$.

And if $n=2m$, then

$y_{(m)} = a + bx_{(m)}$, $y_{(m+1)} = a + bx_{(m+1)}$ if $b > 0$, and
$y_{(m)} = a + bx'_{(m)} = a + bx_{(m+1)}$,
$y_{(m+1)} = a + bx'_{(m+1)} = a + bx_{(m)}$ if $b < 0$.

Hence, in either case, we have $\bar{y} = a + b\bar{x}$.

Finally, if x is discrete, so is y and the value with the highest frequency of y must correspond to the value with the highest frequency of x. If x is continuous, so is y and the frequency density of any y-value, say g(y), is related to the frequency density of the corresponding x-value, say f(x), by

This result will be proved in Unit 10.

$$g(y) = f\left(\frac{y-a}{b}\right)\left|\frac{dy}{dx}\right|^{-1}$$
$$= \frac{1}{|b|} f\left(\frac{y-a}{b}\right).$$

Thus, the frequency density at any value of y is just $\frac{1}{|b|}$ times the frequency density at the corresponding value of x. Consequently, if there is a value $\overset{\circ}{x}$ of x with the highest frequency density, then $a+b\overset{\circ}{x}$ must be the value of y with the highest frequency density.

Hence, $\overset{\circ}{y} = a + b\overset{\circ}{x}$,

whether the variables are discrete or continuous.

iii) If we have ungrouped data, then

$$\sum_1^n x_i = n\bar{x} \Rightarrow \sum_1^n x_i - n\bar{x} = 0 \Rightarrow \sum_1^n (x_i - \bar{x}) = 0.$$

For grouped data, we similarly have

$$\sum_1^k f_i x_i = n\bar{x} \Rightarrow \sum_1^k f_i x_i - \bar{x}\sum_1^k f_i = 0$$

$$\Rightarrow \sum_1^k f_i(x_i - \bar{x}) = 0.$$

iv) Let x_{ij} be the jth observation in the ith set ($i=1,2,\ldots,k$ and $j=1,2,\ldots,n_i$). Then the mean of the ith set, $\bar{x}_i$, is given by

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \quad (i=1, 2,\ldots k) \qquad \ldots (1)$$

As to the grand mean $\bar{x}$, we have

$$\bar{x} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{\sum_1^k n_i}$$

$$= \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{}$$

Now we are listing some exercises for you to solve. You can use the five properties that we have just discussed to solve these.

E6) If two sets of observations are combined, show that the composite mean must lie between the two set means.

E7) Let y be a monotone function of x, say g(x). Show that

a) $\bar{y} = g(\bar{x})$.

b) Show that $\overset{\circ}{y} = g(\overset{\circ}{x})$ if the variables are discrete.

c) Is $\bar{y} = g(\bar{x})$?

(**Hint :** Try with $g(x) = x^3$.)

A monotone function is an increasing or a decreasing function.

E8) Prove algebraic properties i) and v).

E9) The mean of a number of temperature readings on the Centigrade (Celsius) scale is 33.2 degrees. What would be the mean if the readings were taken on the Fahrenheit scale?

E10) There are four blocks in an urban locality, having 126, 153, 137 and 190 households. If the mean income (in rupees) for a month per household is 2012.35, 1972.45, 2734.56 and 2415.67 for the four blocks, respectively, then what is the mean household income for the month for the locality as a whole?

So far we have seen how to compute some measures of central tendency and have also discussed some of their algebraic properties. Now, we should be able to decide which of these measures should be chosen for the given data. For this, we have to know the pros and cons of using each of these measures. In the next sub-section, we'll talk about just this.

2.3.5 A Comparison of the Measures

Can you think of some conditions which a good measure should satisfy? Here we list some.

- It should be rigidly defined and easy to interpret.
- It should not be too difficult to compute.
- It should also be based on all the observations made. At the same time, it should not be too highly affected by comparatively few extreme observations (i.e., observations that are extremely large or extremely small).
- Later, in this block, you will see that we use a measure of central tendency in the computation of various other measures like the standard deviation and the mean deviation. So it should be possible to subject the measure of central tendency to further mathematical treatment.

Now let us compare the mean, median and the mode with respect to each of these criteria.

As you have seen, the mean is rigidly defined, and so is the median, though its definition may not lead to a unique median. The position of the mode is somewhat similar to that of the median.

All these measures are easy to interpret and not too difficult to compute, although, there is no satisfactory method for the determination of the mode from a frequency table in the continuous case.

In the computation of each of the measures, all the observations have to be taken into account. But it is only the mean that directly depends on all the observations: a change in any one of the observations influences the value of the mean but the median and mode are not so sensitive.

As regards amenability to algebraic treatment, too, Section 2.3.4 shows that the mean is the best.

Considering all these facts, we can say that the mean is, generally, the best measure of central tendency. But here is a warning. If the data have even a few observations of an extreme type, these may make the mean unrepresentative of the data. In such

cases, we may have to choose another measure to represent the data. This point will be clear to you after you have done E11).

E11) In a group of 10 male college students, there is an exceptionally tall boy. The following figures (arranged in ascending order) indicate their stature in cm.

161.5 161.6 161.9 162.0 162.3 162.3 165.1 165.4 165.6 186.5

Find the mean stature per student of the group and see if it is a representative value. What is the median? Is that a representative value?

With this, we end our discussion of the measures of central tendency. In the next section, we'll talk about the measures of dispersion.

2.4 MEASURES OF DISPERSION

The commonly used measures of dispersion are the range, the mean (or absolute) deviation and the standard deviation. We start our discussion with the range.

2.4.1 The Range

The range, R, of a variable is the difference between the largest and the smallest observation.

Thus,

$$R = x_{(n)} - x_{(1)} = x'_{(1)} - x'_{(n)}.$$

When the data are in the form of an ungrouped frequency distribution, then we can calculate R exactly. Here

$$R = x_k - x_1.$$

But when the data are presented in the form of a grouped frequency distribution, we can compute R only approximately. In this case, we estimate R by,

$$R = x_{ku} - x_{1l},$$

where x_{ku} is the upper boundary of the last (kth) class and x_{1l} is the lower boundary of the first class.

This estimate is usually an overestimate of the true range.

We can verify this by comparing the range of the raw data on petiole length (Table 5, Unit 1) and that of its frequency distribution given in Table 6, Unit 1. We have

$$x_{1l} = 0.55 \text{ cm}, \quad x_{ku} = 9.35 \text{ cm}.$$

Hence the range would be taken as

$$R = 9.35 - 0.55 = 8.80 \text{ cm}.$$

That it is an overestimate is evident from the raw data which show that $x_{(n)} = 9.2$ cm and $x_{(1)} = 0.8$ cm. Hence, apart from the unavoidable rounding-off errors, the range is exactly

$$9.2 \text{ cm} - 0.8 \text{ cm} = 8.4 \text{ cm}.$$

You have also come across some frequency distributions which have open-ended end classes (see the table in E3 and the adjoining remark). In such situations, it is impossible to find the range. In the following example, we take some sets of data that you have already encountered and find the respective ranges.

Example 8 : i) For the data on daily milk yield for a dairy farm, we have, with n=10,

$$x_{(1)} = 179.5 \text{ litres}, \quad x_{(n)} = 224.3 \text{ litres}.$$

Hence, the range is

$$R = 224.3 - 179.5 = 44.8 \text{ litres}.$$

ii) For the grouped data on household size as shown in Table 3 in Unit 1, we have

$$R = 9 - 1 = 8 \text{ exactly}.$$

Now we turn our attention to the mean deviation.

2.4.2 The Mean Deviation

Suppose A is the chosen average value of x for a given set of observations on the variable. Then we can take the deviations $x_i - A$ $(i=1, 2, \ldots\ldots, n)$ into consideration in constructing a measure of dispersion of x. You would agree that the bigger these deviations are the larger is the dispersion. Would you also agree that the magnitude of the deviations, and not their signs, are important? To get rid of the signs, we consider the absolute values $|x_i - A|$ and get a measure of dispersion by taking their mean, called the **mean deviation** (or mean absolute deviation) of x **about A**. Note that the mean deviation has the same unit in which the original observations are recorded. If we denote this by MD_A, then we have

$$MD_A = \frac{1}{n}\sum_{1}^{n}|x_i - A| \qquad \ldots (13)$$

For data on a variable x, given in the form of an ungrouped frequency distribution, the mean deviation is given exactly by

$$MD_A = \frac{1}{n}\sum_{1}^{k} f_i|x_i - A| \qquad \ldots (14)$$

To compute the mean deviation about A for a grouped frequency distribution, we use the formula

$$MD_A = \frac{1}{n}\sum_{1}^{k} f_i|x_i - A| \qquad \ldots (15)$$

The only difference between Formulas (14) and (15) is that in (14), x_i denotes the ith distinct value of the discrete variable and in (15), x_i denotes the class-mark of the ith class.

The following remark contains an important result.

Remark 2 : The mean deviation MD_A is least when A is the median of x. We are not going to prove this here. But if you are interested you can look up the book : *Fundamentals of Statistics,* Vol. 1 by Goon, Gupta and Dasgupta. This book is available in your study centre library.

The result stated in Remark 2 perhaps supplies a rationale for taking the median as the origin while computing the mean deviation of a set of observations.

Now, if the number of observations, n, is odd, say, $n=2m+1$, then $\tilde{x} = x_{(m+1)}$.

$$\therefore nMD_{\tilde{x}} = \sum_{i=1}^{m}(\tilde{x} - x_{(i)}) + \sum_{i=m+2}^{2m+1}(x_{(i)} - \tilde{x})$$

$$= S_2 - S_1,$$

$$\text{where } S_1 = \sum_{1}^{m} x_{(i)} \text{ and } S_2 = \sum_{m+2}^{2m+1} x_{(i)}$$

If $n = 2m$, then $x_{(m)} \leq \tilde{x} \leq x_{(m+1)}$.

$$\therefore nMD_{\tilde{x}} = \sum_{1}^{m}(\tilde{x} - x_{(i)}) + \sum_{m+1}^{2m}(x_{(i)} - \tilde{x})$$

$$= S_2' - S_1',$$

$$\text{where } S_1' = \sum_{1}^{m} x_{(i)} \text{ and } S_2' = \sum_{i=m+1}^{2m} x_{(i)}.$$

Thus, we have

$nMD_{\tilde{x}}$ = Sum of observations exceeding the median − Sum of observations that are less than the median. ... (16)

Using this formula, we can find the mean deviation about the median without explicitly knowing the value of the median. We now give an example to illustrate the use of the formula.

Example 9 : Consider the data on the daily yield of milk (in litres) of a dairy farm given in Example 1. Arranged in ascending order, the values are

179.5, 184.7, 185.4, 194.4, 199.7, 203.5, 207.3, 213.7, 218.2, 224.3.

The mean deviation about the median (i.e., about any value between 199.7 and 203.5) is given by

$$10MD_{\tilde{x}} = \sum_{1}^{5} x_{(i)} - \sum_{6}^{10} x_{(i)}$$

$$= 1067.0 - 943.7 = 123.3$$

Hence,

$$MD_{\tilde{x}} = 12.3 \text{ litres.}$$

Now, for these data, $\bar{x}$=201.07 litres. That is, $\bar{x}$ also lies between 199.7 and 203.5. Hence, 12.3 litres is also the mean deviation about the mean.

In the next example, we consider data in the form of an ungrouped frequency table.

Example 10 : Let us calculate $MD_{\tilde{x}}$ for the frequency distribution of household size given in Table 3 in Unit 1. Here, n = 80 and $\tilde{x}$=5.

We show the required computations in the following table.

Table 2

x	f_x	$\lvert x - \tilde{x}\rvert$	$\lvert x - \tilde{x}\rvert f_x$
1	3	4	12
2	7	3	21
3	11	2	22
4	14	1	14
5	19	0	0
6	12	1	12
7	8	2	16
8	4	3	12
9	2	4	8
Total	80	—	117

Now

$$MD_x = \frac{1}{n}\Sigma|x-\tilde{x}|f_x$$

$$= \frac{1}{80} \times 117$$

$$= 14.63 \text{ litres.}$$

Next we take the case of data given in the form of a grouped frequency table.

Example 11 : Let us find $MD_{\tilde{x}}$ for the frequency distribution of petiole length given in Table 6 of Unit 1. Here, $\tilde{x}$=5.359 cm, which lies in the class interval 5.35–6.15.

We first form the following table.

Table 3

Class mark x_i	Frequency f_i	$\lvert x_i - \tilde{x}\rvert$	$f_i\lvert x_i - \tilde{x}\rvert$
0.95	2	− 4.409	8.818
1.75	6	− 3.609	21.654
2.55	8	− 2.809	22.472
3.35	10	− 2.009	20.090
4.15	24	− 1.209	29.016
4.95	43	− 0.409	17.587
5.75	52	0.391	20.332
6.55	33	1.191	39.303
7.35	15	1.991	29.865
8.15	4	2.791	11.164
8.95	1	3.591	3.591
Total	198		223.892

Then $MD_{\tilde{x}} = \frac{1}{n}\sum f_i |x_i - \tilde{x}|$

$= \frac{1}{198}(223.892)$

$= 1.131.$

Try to do these exercises now.

E12) Show that in computing $MD_{\tilde{x}}$, it is necessary to consider only the positive deviations (for which $x_i > \tilde{x}$) or only the negative deviations (for which $x_i < \tilde{x}$). Indeed, if the sum of the former is P and that of the latter is Q, show that

$$nMD_{\tilde{x}} = 2P = -2Q.$$

E13) For the age distribution of Indians shown in E3, obtain the mean deviation about median.

In the next sub-section, we'll introduce standard deviation, which is considered to be the best measure of dispersion.

2.4.3 Standard Deviation

Consider the deviations of the observations on variable x from a chosen average A. Instead of taking their absolute values to free them from their signs, we may take their squares. This leads us to an alternative measure of dispersion. The mean of the squares of deviations is called the mean square deviation about A and may be denoted by s_A^2. Thus,

$$s_A^2 = \frac{1}{n}\sum_1^n (x_i - A)^2$$

The positive square-root of this, denoted by s_A, is an alternative measure of dispersion and is called the **root-mean-square deviation** of x about A. Thus,

$$s_A = \sqrt{\frac{1}{n}\sum_1^n (x_i - A)^2}. \qquad \ldots (17)$$

s_A has the same units as x.

For grouped data, we take

$$s_A = \sqrt{\frac{1}{n}\sum_1^k f_i (x_i - A)^2} \qquad \ldots (18)$$

We have the following result, in view of which the root-mean-square deviation is usually measured about the mean.

The root-mean-square deviation is least when $A = \bar{x}$.

Proof : We may, without loss of generality, consider ungrouped data. Then, we have

$$ns_A^2 = \sum_1^n (x_i - A)^2$$

Since $x_i - A = (x_i - \bar{x}) + (\bar{x} - A)$,

We have $(x_i - A)^2 = (x_i - \bar{x})^2 + 2(\bar{x} - A)(x_i - \bar{x}) + (\bar{x} - A)^2$

Thus, $\sum_1^n (x_i - A)^2 = \sum_1^n (x_i - \bar{x})^2 + n(\bar{x} - A)^2.$

$\sum_1^n 2(\bar{x}-A)(x_i-\bar{x})$

$= 2(\bar{x}-A)\sum_1^n (x_i-\bar{x})$

$= 0.$

Since $n(\bar{x} - A)^2 \geq 0$, we can say that

$\sum_1^n (x_i - A)^2$, and hence s_A^2, is least if and only if

$A = \bar{x}$

So, in computing the root-mean-square deviation, we usually take deviations about the mean. The root-mean-square deviation about the mean is called the **standard deviation** and is denoted by s. The square of the standard deviation is known as the **variance.** The unit for standard deviation is the same in which the original observations are recorded, while for variance, it is the square of the same unit.

Thus, $s = \sqrt{\frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})^2}$ ... (19)

Since $(x_i - \bar{x})^2 = x_i^2 - 2\bar{x}\,x_i + \bar{x}^2$,

we get

$$\sum_{1}^{n}(x_i - \bar{x})^2 = \sum_{1}^{n}x_i^2 - 2\bar{x}\sum_{1}^{n}x_i + n\bar{x}^2$$

$$= \sum_{1}^{n}x_i^2 - n\bar{x}^2, \quad \text{since } \sum_{1}^{n}x_i = n\bar{x}.$$

Thus, $s = \sqrt{\frac{1}{n}\sum_{1}^{n}x_i^2 - \bar{x}^2}$ ... (20)

is another expression for s.

For grouped data, we have the formula

$$s = \sqrt{\frac{1}{n}\sum_{1}^{k}f_i(x_i - \bar{x})^2}, \quad \text{... (21)}$$

$$\text{or } s = \sqrt{\frac{1}{n}\sum_{1}^{k}f_i x_i^2 - \bar{x}^2}. \quad \text{... (22)}$$

Now we'll illustrate the method of finding the standard deviation of the given data on a continuous variable. We are sure you will have no difficulty in computing the standard deviation of the data on a discrete variable.

Example 12 : For the grouped data on petiole length (see Table 6, Unit 1), the computations needed for determining the standard deviation (together with the mean) may be laid out in tabular form. But, in this case, it will be convenient to subject the variable to a change of base and scale. Let us take $u = \frac{(x - 4.95)}{0.8}$.

Table 4

Class mark x_i	$u_i = (x_i - 4.95)/0.8$	Frequency f_i	$f_i u_i$	$f_i u_i^2$
0.95	−5	2	−10	50
1.75	−4	6	−24	96
2.55	−3	8	−24	72
3.35	−2	10	−20	40
4.15	−1	24	−24	24
4.95	0	43	0	0
5.75	1	52	52	52
6.55	2	33	66	132
7.35	3	15	45	135
8.15	4	4	16	64
8.95	5	1	5	25
Total	—	198	82	690

We have, from this table,

$$\sum_{1}^{k}f_i u_i = 82$$

$$\sum_{1}^{k}f_i u_i^2 = 690.$$

Hence $ns_u^2 = \sum_1^k f_i u_i^2 - \frac{(\sum_1^k f_i u_i)^2}{n}$

$$= 690 - \frac{(82)^2}{198}$$
$$= 690 - 33.9596$$
$$= 656.0404$$

Hence, the variance of u is

$$s_u^2 = \frac{656.0404}{198} = 3.3133 \text{ (cm)}^2$$

But how can we get s^2, the variance of x, from this? It is related to the variance of u by

$$s^2 = c^2 s_u^2.$$

Here c = 0.8. Therefore, we have

We'll prove this in the next section.

$$s^2 = 0.64 \times 3.3133 = 2.1205 \text{ (cm)}^2,$$

and the standard deviation is

$$s = \sqrt{2.1205} = 1.456 \text{ cm.}$$

From the same table, we get

$$\bar{u} = \frac{82}{198} = 0.4141 \text{ cm,}$$

so that

$$\bar{x} = 4.95 + 0.8 \times 0.4141$$
$$= 4.95 + 0.331 = 5.281 \text{ cm.}$$

Check if you can find the standard deviation of the data on a discrete variable by solving this exercise now.

E14) Find the standard deviation of the data
   a) on the daily output of milk given in Example 1.
   b) on the size of households given in Table 3 in Unit 1.

Next, we shall discuss some algebraic properties of the three measures of dispersion discussed so far.

2.4.4 Algebraic Properties of the Measures

The range, the mean deviation about A (which may be taken to be either the mean or the median or the mode) and the standard deviation have the following algebraic properties. We will prove some and you can prove the rest.

1) If all the observations on the variable x are equal, say to a, then $R = MD_A = s = 0$. This is very easy to prove and we are sure you will be able to prove it (see E 15).

2) Let $y = a + bx$ $(b \neq 0)$ and R_x, MD_A^x and s_x be the range, mean deviation about A (an average value of x) and standard deviation of x, while R_y, MD_A^y, and s_y be the corresponding quantities for y. Then

   i) $R_y = |b| R_x$,

   ii) $MD_A^y = |b| MD_A^x$,

   iii) $s_y = |b| s_x$.

Proof : i) **Range :** If $b > 0$, then

$$y_{(1)} = a + bx_{(1)}, \; y_{(n)} = a + bx_{(n)},$$

so that

$$R_y = y_{(n)} - y_{(1)} = b(x_{(n)} - x_{(1)}) = bR_x.$$

If $b < 0$, then

$$y_{(1)} = a + bx_{(n)}, \; y_{(n)} = a + bx_{(1)},$$

so that

$$R_y = y_{(n)} - y_{(1)} = b(x_{(1)} - x_{(n)}) = -bR_x.$$

Hence, in either case,

$$R_y = |b|R_x.$$

ii) **Mean deviation :** We have already seen (in Section 2.3.4) that under a transformation of this type,

$$A' = a + bA.$$

Now, $y - A' = (a + bx) - (a + bA)$
$= b(x - A).$

Hence

$$MD_A^y = \frac{1}{n}\sum_1^n |y_i - A|$$

$$= |b| \times \frac{1}{n}\sum_1^n |x_i - A|$$

$$= |b|MD_A^x .$$

iii) We leave this to you as an exercise (see E15) b).

Remark 3 : Note that a good measure of dispersion should have Property (2). For, if the observations are all increased or decreased by a constant amount, then the dispersion remains unchanged, but if they are all increased or decreased in a constant proportion, then the dispersion too gets increased or decreased in the same proportion.

3) Suppose several (say k) sets of observations on x are combined, the ith set having n_i observations with mean $\bar{x}_i$ and standard deviation s_i (i=1, 2,.....,k). Then the composite standard deviation, i.e., the standard deviation of the combined data is s, given by

$$s^2 = \frac{\sum_1^k n_i s_i^2 + \sum_1^k n_i (\bar{x}_i - \bar{x})^2}{\sum_1^k n_i}$$

Proof : Let x_{ij} be the jth observation in the ith set (i=1,2,...,k and j = 1,2,....., n_i). Then

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n_i} x_{ij} , \quad s_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

and, as we have seen in Section 2.3.4,

$$\bar{x} = \frac{\sum_1^k n_i \bar{x}_i}{\sum_1^k n_i}$$

Now, s^2 is the variance of the combined set and is, by definition, given by

$$\left(\sum_1^k n_i\right) s^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2.$$

But $x_{ij} - \bar{x} = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})$,

so that $(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_i)^2 + 2(\bar{x}_i - \bar{x})(x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})^2$,

and $\sum_j (x_{ij} - \bar{x})^2 = \sum_j (x_{ij} - \bar{x}_i)^2 + n_i(\bar{x}_i - \bar{x})^2$, since $\sum_j (x_{ij} - \bar{x}_i) = 0$.

$$= n_i s_i^2 + n_i(\bar{x}_i - \bar{x})^2$$

Hence,

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i n_i s_i^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2 ,$$

and the result is established.

Remark 4 : This result shows that the standard deviation of the combined set may be non-zero even when the individual sets have zero standard deviations. If the values in the ith set are all equal to a_i, i=1,2,.....,k,

and $\bar{a} = \dfrac{\sum_1^k n_i a_i}{\sum_1^k n_i}$, then s^2 will equal $\dfrac{\sum n_i (a_i - \bar{a})^2}{\sum n_i}$.

This will be non-zero unless $a_1 = a_2 = ... = a_k$.

Try this exercise now.

E15) Prove a) Property (1) and
b) Property (2) (iii).

The next example shows how these properties are useful in some real-life situations.

Example 13 : For the four blocks in an urban locality, the means and standard deviations of household income (in rupees) for a certain month are given below together with the number of households in each block:

Block	I	II	III	IV
Number of households	126	153	137	190
Mean Income (in Rs.)	2012.35	1972.45	2734.56	2415.47
S.d of income (in Rs.)	153.17	189.62	183.47	202.09

Let us find the mean and standard deviation for the entire locality.

We have, in the notation used in (3),

$$\sum_1^k n_i = 606, \quad \sum_1^k n_i \bar{x}_i = 1388914.97$$

and $$\sum_1^k n_i s_i^2 = 20828581.0$$

Also, $\bar{x} = \sum_1^k n_i \bar{x}_i \Big/ \sum_1^k n_i = 2291.94$ (rupees).

We show the necessary computations in the table below :

Table 5

n_i	$\bar{x}_i$	s_i	$n_i\bar{x}_i$	$n_i s_i^2$	$(\bar{x}_i - \bar{x})$	$n_i(\bar{x}_i - \bar{x})^2$
126	2012.35	153.17	253556.70	2956092.2	−279.59	9849491.6
153	1972.45	189.62	301784.85	5501228.9	−319.49	5617301.0
137	2734.56	183.47	374634.72	4611590.0	442.62	26840008.0
190	2415.47	202.09	458939.30	7759669.9	123.53	7759669.9
606	—	—	1388914.97	20828581.0	—	60066470.5

Hence the composite variance is

$$s^2 = \frac{20828581.0 + 60066470.5}{606} 133490.18$$

and the composite standard deviation is

$$s = \sqrt{133490.18} = 365.36 \text{ rupees.}$$

See if you can solve these exercises now.

E16) If R and s be the range and the standard deviation for a set of n observations on a variable x, show that

$$\frac{R^2}{2n} \leq s^2 \leq \frac{R^2}{4}.$$

E17) Show that the standard deviation can be expressed in terms of the mutual differences $x_i - x_j$ of the observations – more precisely,

$$s^2 = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2.$$

E18) There are 4 sections, A,B,C and D in Class X of a school, having 48, 41, 52 and 45 students, respectively. If the mean IQs per student for the sections are 133.2, 125.4, 110.5 and 97.8, and the standard deviations of IQ are 3.8, 4.7, 5.1 and 5.9, respectively, then find the composite mean and composite standard deviation of IQ for the class.

So far, in this section, we have discussed three measures of dispersion. Now let us compare and contrast these.

2.4.5 A Comparison of the Measures of Dispersion

In Sec. 2.3.5, we discussed the criteria which a measure of central tendency should fulfil. Now, can you think of some criteria which a good measure of dispersion should satisfy? Actually all the criteria for a measure of central tendency apply to a measure of dispersion too. In addition, a measure of dispersion, to be considered satisfactory, should properly reflect the variability of the observations on the given variable. Before we go any further, why don't you go back to Section 2.3.5, and take a quick look at the criteria listed there?

Now let us compare our three measures of dispersion in the light of these criteria.

- The range and mean deviation are easy to interpret, while the standard deviation may seem a little difficult.
- The range is generally easy to determine, while the determination of the mean deviation and standard deviation is more difficult.
- The range depends only on the two extreme observations and so ignores the variation of the data lying in between. As such, two or more distributions with widely different degrees of dispersion may have the same range. The mean and standard deviation, on the other hand, properly reflect the variation in the data.
- For the same reason, compared to the mean deviation or the standard deviation, the range is highly affected by the presence of an extremely high or extremely low value.
- As regards amenability to algebraic treatment, the standard deviation is surely the best of the three measures.

We now summarise these observations in the following table. We have deliberately left some gaps and expect you to fill them (see E19).

Table 6 : Comparison of the measures of dispersion

Range	Mean Deviation	Standard Deviation
Easy to compute	—	—
—	Properly reflects the variation in the data	—
—	—	Not unduly affected by the presence of extremely high or low values.
Not amenable to algebraic treatment	—	—

E19) Fill in the gaps in Table 6.

In the light of these observations, the range appears to be the worst measure. Indeed, the standard deviation may generally be taken to be the best measure of dispersion, just as the mean may generally be taken to be the best measure of central tendency. However, in industrial applications, to check whether a manufacturing process is under control, we have to compute a measure of dispersion of some important characteristic of the manufactured product at frequent intervals. Due to the simplicity of calculation, the range is quite popular in such cases.

In the last two sections, we have discussed the measures of central tendency and dispersion. In the next section, we'll see how we can compare two or more data sets.

2.5 COEFFICIENT OF VARIATION

Sometimes we need to compare two or more distributions in respect of dispersion. Then it becomes necessary to relate a usual measure of dispersion to a measure of central tendency.

Suppose a height distribution is to be compared with a weight distribution in respect of dispersion. Since the height data will be in, say, cm, while the weight data will be in, say, kg, a comparison of the two standard deviations (or even the two mean deviations or ranges) will not make sense. This is because these measures will have different units attached to them. To overcome this difficulty, we take

$$v = \frac{100\ s}{\bar{x}}, \qquad \ldots \quad (23)$$

(where $\bar{x}$ is assumed to be non-zero) which is free of units, i.e., is only a number. This measure will be more appropriate for comparisons. The measure (23), that expresses the standard deviation of x as a percentage of the mean, is called the **coefficient of variation** (CV) of x.

Such a measure becomes useful even when several sets of data expressed in the same units but with widely differen averages are to be compared. For example, consider the following situation. Suppose Firm I manufactures ball-bearings meant for bicycles, while Firm II manufactures ball-bearings that are to be used in motor cars. Naturally, the ball bearings of Firm I have to be much smaller in volume than those of Firm II. If we want to know which of the two firms produces ball-bearings with a lesser variation in size, then a comparison of the two standard deviations will not be relevant. Since the ball-bearings produced by Firm II are larger, we would be ready to tolerate higher deviation from the mean size (and hence a higher standard deviation) than in the case of Firm I. A comparison of the two coefficients of variation will be much more meaningful. For the same reason, we should use the coefficient of variation, rather than the standard deviation, when income disparities in, say, a group of managers are to be compared with those in a group of clerks.

Example 14 : The mean and standard deviation of family income (in US dollars) in a year are given below for each of the three countries, A, B and C.

Country	A	B	C
Mean	18,727	320	339
Standard deviation	2,432	54	21

We would like to know which of the three countries shows the highest disparity in family income and which one shows the lowest.

We may compare the standard deviations of B and C, but A surely stands in a different category, having a mean family income that is vastly higher than those of B. The comparison should, therefore, be made in terms of the coefficients of variation rather than the standard deviations. We have

$$v_A = 100 \times \frac{2432}{18727} = 12.99\% ,$$

$$v_B = 100 \times \frac{54}{320} = 16.88\% ,$$

$$v_C = 100 \times \frac{21}{339} = 6.19\% ,$$

Hence, income disparities are the highest in B and the lowest in C, while A stands in between B and C.

Would you like to try this exercise now?

E20) The following values are for a group of male undergraduates :

Mean height—159.8 cm, Mean weight—50.27 kg.
S.d. of height—11.3 cm, S.d. of weight—4.74 kg.

Is it correct to say that their weights show greater variation than their heights? Why?

With this, we bring this unit to a close. Here is a brief summary of our discussion.

2.6 SUMMARY

In this unit, we have discussed some features of univariate distributions. In particular, we have seen that

1) the observations on a variable show a tendency to cluster around some point or a small part of the range of variation. This is called the **central tendency.** An **average** is a value which can be taken to be representative of the data. The variation of the observations from the average is called **dispersion.**

2) there are different types of measures of central tendency:

   the **mean,** the **median** and the **mode.**

   We have seen how to compute these from raw data, grouped or ungrouped frequency distributions.

   We have noted that the mean, median and mode have certain algebraic properties.

   We have also discussed the relative advantages and disadvantages of these measures.

3) there are various measures of dispersion :
   the **range,** the **mean deviation,** the **standard deviation.**

   We have noted the algebraic properties and relative merits and demerits of these measures.

4) **coefficient of variation** is used to compare dispersions of different data sets.

2.7 SOLUTIONS AND ANSWERS

E1) a) Data : 75, 75, 80, 83, 105

$\therefore x_{(3)} = 80$

$x'_{(2)} = 83$

b) $F_i + F'_{i+1} = \sum_{1}^{i} f_j + \sum_{i+1}^{k} f_j = \sum_{1}^{k} f_j = n$

$\Rightarrow F_i = n - F'_{i+1}$

Similarly, $F'_i + F_{i-1} = n$ and hence

$F'_i = n - F_{i-1}$

c) $F_k = F'_1 = n.$

E2) mean score $= \frac{838}{15} = 55.86.....$

E3)

x_i	f_i	u_i	f_iu_i	F_i
2.5	12.59	−6	−75.54	12.59
7.5	14.08	−5	−70.4	26.67
12.5	12.88	−4	−51.52	39.55
17.5	9.63	−3	−28.89	49.18
22.5	8.62	−2	−17.24	57.8
27.5	7.63	−1	− 7.63	65.43
32.5	6.38	0	0	71.81
37.5	5.85	1	5.85	77.66
42.5	5.14	2	10.28	82.8
47.5	4.40	3	13.2	87.2
52.5	3.83	4	15.32	91.03
57.5	2.47	5	12.35	93.5
67.5	6.49	7	45.43	100
	100		−148.79	

$$\bar{u} = \frac{\sum f_i u_i}{\sum f_i} = \frac{-148.79}{100} = -1.4879$$

$\therefore \bar{x} = 5\bar{u} + 32.5$

$= -7.4395 + 32.5$

$= 25.06$ years.

E4) $x_l = 20, F_l = 49.18, n = 100, c = 5, f_0 = 8.62.$

$$\therefore \tilde{x} = x_l + \frac{(n/2) - F_l}{f_0} \times c$$

$= 20.47$ years.

E5) a) $x_l = 5, f_0 = 14.08, f_- = 12.59, f_+ = 12.88, c = 5.$

$$\therefore \overset{\circ}{x} = x_l + \frac{f_0 - f_-}{2f_0 - f_- - f_+} \times c$$

$= 7.77$ years.

b) Now from E3 and E4, $\bar{x} - \overset{\circ}{x} = 17.29$ years
and $\bar{x} - \tilde{x} = 4.59$ years.
So the empirical relation is not borne out in this case.

E6) Let $\bar{x}_1$ be the mean and n_1 be the total frequency of the first set.

Let $\bar{x}_2$ be the mean and n_2 be the total frequency of the second set.

Let $\bar{x}$ be the composite mean.

Suppose $\bar{x}_1 < \bar{x}_2$, $\bar{x}$

=

...diemost value of x corresponds to the middlemost

b) If x is discrete, so is g(x) and the frequency of any value of x is also the frequency for the corresponding value of y. Hence, the value of y with the highest frequency corresponds to the value of x with the highest frequency, i.e., $\overset{\circ}{y} = g(\overset{\circ}{x})$.

c) $\bar{y} \neq g(\bar{x})$ unless g is a linear function.

E8) i) $x_i = a \ \forall i$

$$\Rightarrow \bar{x} = \frac{1}{n}\sum_{1}^{n} x_i = a$$

The median and the mode are, obviously, equal to a.

ii) $z_i = x_i + y_i$

$$\Rightarrow \bar{z} = \frac{1}{n}\sum_{1}^{n} z_i$$

$$= \frac{1}{n}\sum (x_i + y_i)$$

$$= \frac{1}{n}\sum x_i + \frac{1}{n}\sum y_i$$

$$= \bar{x} + \bar{y}.$$

Similar proof for the case when $z_i = x_i - y_i$.

E9) $32 + \frac{9}{5} \times 33.2 = 91.76°$

E10) $$\bar{\bar{x}} = \frac{2012.35 \times 126 + 1972.45 \times 153 + 2734.56 \times 137 + 2415.67 \times 190}{126 + 153 + 137 + 190}$$

$$= \frac{1388952.97}{606} = \text{Rs. } 2292.$$

E11) $\bar{x} = 165.42$ cm. $\tilde{x} = 162.3$ cm.
165.42 is not a representative value.
162.3 is one.

E12) $P + Q = \sum_{1}^{n} (x_i - \bar{x}) = 0 \Rightarrow P = -Q.$

$nMD_{\bar{x}} = P - Q = 2P = -2Q.$

E13)

$\tilde{x} = 20.48$

x_i	f_i	$f_i\lvert x_i - \tilde{x}\rvert$
2.5	12.59	226.36
7.5	14.00	182.75
12.5		102.78
17.5		28.69
22.5	8.6	17.41
27.5	7.63	53.56
32.5	6.38	58
37.5	5.85	
42.5	5.14	
47.5	4.40	
52.5	3.83	
57.5	2.47	
67.5	6.49	
	100	

$A = a$

$\therefore MD_A = \frac{1}{n}\sum|x_i - a| = \frac{1}{n}\sum|a-a| = 0.$

Similarly $s = 0$.

b) $y - \bar{y} = b(x_i - \bar{x})$, since $\bar{y} = a + b\bar{x}$.

$\therefore s_y^2 = \frac{1}{n}\sum_1^n (y_i - \bar{y})^2$

$= b^2 \cdot \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$

$= b^2 s_x^2$

$\therefore s_y = |b| s_x$, since standard deviation is the positive square root of variance.

$\therefore MD\bar{x} = \frac{1539.54}{100} = 15.39$

E14)

a)

x_i	x_i^2
218.2	47611.24
199.7	39880.09
207.3	42973.29
185.4	34373.16
213.7	45667.69
184.7	34114.09
179.5	32220.25
194.4	37791.36
224.3	50310.49
203.5	41412.25
	406353.91

$\bar{x} = 201.07$

$s^2 = \frac{f}{n}\sum x_i^2 - \bar{x}^2$

$= 40635.391 - 40429.14$

$= 206.25$

$s = 14.36$ litres

b)

x_i	f_i	$f_i x_i$	$f_i x_i^2$
1	3	3	3
2	7	14	28
3	11	33	99
4	14	56	224
5	19	95	475
6	12	72	432
7	8	56	392
8	4	32	256
9	2	18	162
	80	379	2071

$\bar{x} = 4.737$

$s^2 = \frac{1}{n}\sum f_i x_i^2 - \bar{x}^2$

$\simeq 25.8875 - 22.4439$

$= 3.44$

$\therefore s = 1.86$

E15) a) $x_i = a \quad \forall_i$.

$\Rightarrow R = x_{(n)} - x_{(1)} = a - a = 0.$

E16) $ns^2 = \sum_1^n (x_{(i)} - \bar{x})^2 \geq (x_{(1)} - \bar{x})^2 + (x_{(n)} - \bar{x})^2$

$$\geq \left\{x_{(1)} - \frac{x_{(1)} + x_{(n)}}{2}\right\}^2 + \left\{x_{(n)} - \frac{x_{(1)} + x_{(n)}}{2}\right\}^2$$

$= R^2/2$

and

$$ns^2 = \sum_1^n (x_{(i)} - \bar{x})^2 \leq \sum_{x_{(i)} \leq a} (x_{(i)} - a)^2 + \sum_{x_{(i)} > a} (x_{(i)} - a)^2 ,$$

where $a = \frac{\{x_{(1)} + x_{(n)}\}}{2}$.

Again,

$$\sum_{x_{(i)} \leq a} (x_{(i)} - a)^2 \leq n_1 \{x_{(1)} - a\}^2$$

$$= \frac{n_1}{4} R^2$$

and

$$\sum_{x_{(i)} > a} (x_{(i)} - a)^2 < n_2 \{x_{(n)} - a\}^2$$

$$= \frac{n_2}{4} R^2 .$$

Since $n_1 + n_2 = n$, the inequality $s^2 \leq R^2/4$ follows.

E17) Write $x_i - x_j = (x_i - \bar{x}) - (x_j - \bar{x})$, so that

$(x_i - x_j)^2 = (x_i - \bar{x})^2 - 2(x_i - \bar{x})(x_j - \bar{x}) + (x_j - \bar{x})^2$.

Sum with respect to both i and j and note that

$\sum_i (x_i - \bar{x}) = \sum_j (x_j - \bar{x}) = 0$ and $\sum_i (x_i - \bar{x})^2 = \sum_j (x_j - \bar{x})^2 = ns^2$.

E18) $s = 21.3$, $\bar{x} = 116.57$.

E19)

Range	Mean Deviation	Standard Deviation
Easy to compute	More difficult	More difficult
Does not reflect variation properly	Properly reflects variation	Properly reflects variation
Greatly affected by the presence of extreme values	Not unduly affected	Not unduly affected
Not amenable to algebraic treatment	Not easily amenable to algebraic treatment	Is amenable to algebraic treatment

E20) Yes, since the CVs are 7.07% and 9.43%, respectively.

UNIT 3 SKEWNESS AND KURTOSIS

Structure

3.1 INTRODUCTION

In Unit 2, we talked about the central tendency and dispersion of frequency distributions. We have also seen how to compute some measures of central tendency and dispersion. Now, in this unit, we shall discuss two additional features of frequency distributions. These are : **skewness** and **kurtosis**. A measure of skewness would tell us how far the frequency curve of the given frequency distribution deviates from a symmetric one. On the other hand, a measure of kurtosis gives us some information about the degree of flatness (or peakedness) of the frequency curve. So, these two features, along with the two discussed in the previous unit should give us a good idea about the given frequency distribution.

The measures of skewness and kurtosis that we are going to discuss here, make use of **moments** and **quantiles.** So, we shall first introduce these in Section 3.2.

While studying this unit, you will need to look back at the tables of data given in Unit 1. You will also need your calculator. Check all the calculations in the solved examples, so that you don't have any difficulty in solving the exercises later.

With this unit, we end our discussion of the descriptive measures of univariate data. In the next unit, Unit 4, we'll talk about bivariate data, i.e., data concerning two variables.

Objectives

After reading this unit, you should be able to :

- calculate the moments and the quantiles of a given frequency distribution,
- compute some measures of skewness and kurtosis,
- discuss the relative advantages and disadvantages of these measures.

3.2 MOMENTS AND QUANTILES

As we have mentioned in the Introduction, using moments and quantiles, we can define some measures of skewness and kurtosis. So we can say that moments and quantiles give us some information about the nature of a given frequency distribution. You will see that the mean of a frequency distribution can also be considered as its moment, whereas the median can be considered as its quantile. Now let's study these one by one.

3.2.1 Moments of a Frequency Distribution

If you have studied a little physics, you would have come across the word, "moment". Moment, in physics, measures the tendency of a force to produce rotation. If there

are n forces, $f_1, f_2, \ldots\ldots, f_n$ acting at distances $x_1, x_2, \ldots\ldots, x_n$ from the origin, then the moment of the total force is

$$\frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \qquad \ldots (1)$$

Now isn't this a familiar expression? If in (1), we take f_i to be the frequency of x_i, $i=1, 2,\ldots\ldots, n$, then (1) also gives us the mean of the distribution of x-values. It is because of this similarity, that the term, "moment" has found its way into Statistics. Now let's try to understand this term in the context of Statistics.

You know that the mean and the variance of data on a variable x are given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i = \frac{1}{n}\sum_{i=1}^{k} f_i (x_i - 0), \text{ and}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2, \text{ respectively.}$$

Now, taking a cue from this, if A is any number, we define the **rth moment of x about A** to be the mean of the rth power of the deviations of x from A. We denote this by $m'_r(A)$, or simply m'_r, if there is no confusion about the origin chosen. Thus,

$$m'_r(A) = \frac{1}{n}\sum_{i=1}^{k} f_i (x_i - A)^r \qquad \ldots (2)$$

For raw data, we can write $m'_r(A) = \frac{1}{n}\sum_{1}^{n} (x_i - A)^r$.

If we take $A = \bar{x}$, then we get what are called the **central moments.** The rth central moment is denoted by m_r. Thus

$$m_r = \frac{1}{n}\sum_{i=1}^{k} f_i (x_i - \bar{x})^r \qquad \ldots (3)$$

For raw data, $m_r = \frac{1}{n}\sum_{1}^{n} (x_i - \bar{x})^r$.

We are sure you would agree with the following :

$$m'_0 = m_0 = 1 \qquad \ldots (4)$$

$$m'_1(0) = \bar{x} \qquad \ldots (5)$$

$$m_1 = 0 \qquad \ldots (6)$$

$$m_2 = s^2 \qquad \ldots (7)$$

(5) and (7) say that the mean of a variable is its first moment about zero, while the variance is its second central moment. Recall that we have already established (6) in Section 2.3.4.

Now let us try to establish a connection between central moments and moments about any value A. For simplicity we consider raw data only.

For every i, we have

$$\begin{aligned} x_i - \bar{x} &= (x_i - A) - (\bar{x} - A) \\ &= (x_i - A) - \frac{1}{n}\sum (x_i - A) \\ &= (x_i - A) - m'_1 \end{aligned}$$

Therefore, using binomial theorem, we get

$$(x_i - \bar{x})^r = (x_i - A)^r - \binom{r}{1}(x_i - A)^{r-1} m_1' + \binom{r}{2}(x_i - A)^{r-2}(m_1')^2$$

$$- \ldots\ldots\ldots + (-1)^{r-1}\binom{r}{r-1}(x_i - A)(m_1')^{r-1} + (-1)^r\binom{r}{r}(m_1')^r .$$

If we sum both sides over all i, i=1,2,.....,n, and divide the result by n, we get

$$m_r = m_r' - \binom{r}{1} m_{r-1}' m_1' + \binom{r}{2} m_{r-2}'(m_1')^2 - \ldots\ldots\ldots +$$

$$(-1)^{r-1}\binom{r}{r-1}(m_1')^r + (-1)^r\binom{r}{r}(m_1')^r \quad \ldots (8)$$

Now let us take r = 1. In this case, (8) becomes

$$m_1 = m_1' - \binom{1}{1} m_1' = 0.$$

We have already stated this in (6). Now, if we put r = 2 in (8), we get

$$m_2 = m_2' - \binom{2}{1}(m_1')^2 + \binom{2}{2}(m_1')^2 .$$

This gives us

$$m_2 = m_2' - (m_1')^2 \quad \ldots (9)$$

Putting r = 3 in (8) gives us

$$m_3 = m_3' - 3m_2' m_1' + 2(m_1')^3 \quad \ldots (10)$$

Check that if you put r=4 in (8), you get

$$m_4 = m_4' - 4m_3' m_1' + 6m_2'(m_1')^2 - 3(m_1')^4 \quad \ldots (11)$$

Further, we have $\bar{x} = A + m_1'$ ... (12)

We can use these relationships between the central moments and the moments about any A, for simplifying the calculations involved in the computation of central moments. Here is an example to show how this is done.

We choose A to be a value near the centre of the range of values of x.

Example 1 : Let's evaluate the mean and central moments of the milk yield data (in litres) of a dairy farm first used in Example 1 in Unit 2. Let us take A = 200 litres and first obtain the moments about A. The values of $u = x - A$ and of the squares, cubes and fourth powers of u are shown in the table below. The last column is taken to serve as a check on the calculations, as you will see later.

u_i	u_i^2	u_i^3	u_i^4	$(u_i+1)^4$
18.2	331.24	6028.568	109719.94	135895.45
−0.3	0.09	−0.027	0.01	0.24
7.3	53.29	389.017	2839.82	4745.83
−14.6	213.16	−3112.136	45437.19	34210.20
13.7	187.69	2571.353	35227.54	46694.89
−15.3	234.09	−3581.577	54798.13	41816.16
−20.5	420.25	−8615.125	176610.06	144590.06
−5.6	31.36	−175.616	983.45	447.75
24.3	590.49	14348.907	348678.44	409715.21
3.5	12.25	42.875	150.06	410.06
Total 10.7	2073.91	7896.239	774444.64	818525.85

The last row of the table shows that

$$n=10,$$

$$\sum_1^n u_i = 10.7$$

$$\sum_1^n u_i^2 = 2073.91$$

$$\sum_1^n u_i^3 = 7896.239$$

$$\sum_1^n u_i^4 = 774444.64.$$

Now to check these calculations, we make use of the last column. We have

$$\sum_1^n (u_i + 1)^4 = 818525.85$$

But $\sum_1^n (u_i + 1)^4$ also equals $\sum_1^n u_i^4 + 4\sum_1^n u_i^3 + 6\sum_1^n u_i^2 + 4\sum_1^n u_i + n$

$$= 774444.64 + 31584.96 + 12443.46 + 42.8 + 10$$
$$= 818525.86 .$$

Since we get the same value for $\sum_1^n (u_i + 1)^4$ by these two different methods, we can be sure that the computations are correct.

Now

$$m_1' = \frac{10.7}{10} = 1.07,$$
$$m_2' = 207.391,$$
$$m_3' = 789.624,$$
$$m_4' = 77444.46.$$

Hence

$$\bar{x} = 200 + m_1' = 201.07 \text{ litres, (from (12))}\cdot$$
$$m_2 = m_2' - (m_1')^2 \quad \text{(from (9))}\cdot$$
$$= 207.391 - 1.145 = 206.246 \text{ (litre)}^2,$$
$$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3 \quad \text{(from (10))}$$
$$= 789.624 - 3 \times 221.908 + 2 \times 1.225$$
$$= 792.074 - 665.724 = 726.35 \text{ (litre)}^3\text{s}$$

and
$$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 - 3(m_1')^4 \text{ (from (11))}$$
$$= 77444.46 - 4 \times 844.90 + 6 \times 237.44 - 3 \times 1.31$$
$$= 78869.10 - 3383.53 = 75485.6 \text{ (litre)}^4\text{s}.$$

Try to do this exercise now. The result in this exercise also helps to simplify the computation of central moments.

E1) Suppose the data are subjected to a change of both origin and scale, i.e., let

$$u = \frac{(x-A)}{c} \ (c \neq 0).$$

If $\nu_r' = \frac{1}{n}\sum_1^n u_i^r$, show that

$$\bar{x} = a + c\nu_1' , \ m_2 = c^2 [\nu_2' - (\nu_1')^2]$$

$$m_3 = c^3[\nu_3' - 3\nu_2'\nu_1' + 2(\nu_1')^3] , \ m_4 = c^4[\nu_4' - 4\nu_3'\nu_1' + 6\nu_2'(\nu_1')^2 - 3(\nu_1')^4]$$

Notice how we have used this result in simplifying the computation of the mean and central moments in our next example.

Example 2 : For the frequency table of petiole length of leaves of a pipal tree, let us take $u = (x - 4.95)/0.8$. The table below shows the steps to be followed in obtaining

ν_1', ν_2', ν_3' and ν_4'. Here, again, we take the last column to provide a check on the computations.

u_i	f_iu_i	$f_iu_i^2$	$f_iu_i^3$	$f_iu_i^4$	$f_i(u_i+1)^4$
−5	−10	50	−250	1250	512
−4	−24	96	−384	1536	486
−3	−24	72	−216	648	128
−2	−20	40	−80	160	10
−1	−24	24	−24	24	0
0	0	0	0	0	43
1	52	52	52	52	832
2	66	132	264	528	2673
3	45	135	405	1251	3840
4	16	64	256	1024	2500
5	5	25	125	625	129
	82	690	148	7062	12320

We have $\sum_1^k f_i (u_i + 1)^4 = 12320$

and also $\sum_1^k f_iu_i^4 + 4 \sum_1^k f_iu_i^3 + 6 \sum_1^k f_iu_i^2 + 4 \sum_1^k f_iu_i + n$

$= 7062 + 592 + 4140 + 328 + 198$

$= 12320.$

Hence, we are sure that the column totals are free of errors.

We now have

$\nu_1' = 82/198 = 0.41414$

$\nu_2' = 690/198 = 3.4848,$

$\nu_3' = 148/198 = 0.74747$

$\nu_4' = 7062/198 = 35.667.$

The mean and central moments of petiole length are

$\bar{x} = 4.95 + 0.8 \times 0.41414 = 5.2813$ cm,

$m_2 = (0.8)^2 [3.4848 - (0.41414)^2]$

$= 0.64 \times 3.3133 = 1.7892 \text{ (cm)}^2$

$m_3 = (0.8)^3 [0.74747 - 3 \times 3.4848 \times 0.41414 + 2 \times (0.41414)^3]$

$= 0.512 \times (-3.44006) = -1.7613 \text{ (cm)}^3,$

$m_4 = (0.8)^4 [35.667 - 4 \times 0.74747 \times 0.41414 + 6 \times 3.4848 \times (0.41414)^2 - 3 \times (0.41414)^4]$

$= 0.4096 \times 37.927 = 15.535 \text{ (cm)}^4.$

We may also encounter an exactly opposite situation, that is, given the mean and central moments of a variable x, we may like to express the moments about some other origin, say A, in terms of these quantities. So, we would like to have some formulas which express each m_r' in terms of the central moments. We believe you are now in a position to derive the required formulas.

E2) Prove that $m_r' = m_r + \binom{r}{1} m_{r-1} d + \binom{r}{2} m_{r-2} d^2 + \ldots\ldots\ldots\ldots$

$\ldots\ldots + \binom{r}{r-2} m_2 d^{r-2} + \binom{r}{r} d^r,$

where $d = \bar{x} - A$.

Hint : Use the fact, $x_i - A = (x_i - \bar{x}) + (\bar{x} - A)$.

E3) In Sections 2.3.4 and 2.4.4, you have seen formulas for the composite mean and composite variance in terms of the group means and group variances, when several groups of data on a variable are taken together. Obtain similar formulas for the composite third and fourth central moments, using E 2.

Let us now turn our attention to quantiles.

3.2.2 Quantiles of a Frequency Distribution

Remember how we define median of a given set of data? It is a value x, such that at most **half** of the observations are below x, and at most half are above x. Now, instead of half, if we take a proportion p, $0 < p < 1$, then we get a p-quantile.

Thus, by the **p-quantile** (or **p-fractile**, or **quantile of order p,**) of a variable x, we mean a value, say z_p, of the variable such that at most a proportion p of the observations is below z_p and at most a proportion $(1-p)$ is above z_p. So, the median is the quantile of order $\frac{1}{2}$. With $p = \frac{1}{4}, \frac{1}{2}$ and $\frac{3}{4}$, we have the three **quartiles** $z_{1/4}$, $z_{1/2}$ and $z_{3/4}$ (which are also denoted by q_1, q_2 and q_3). Taking $p = 0.1, 0.2, \ldots\ldots, 0.9$, we get the nine **deciles** and with $p = 0.01, 0.02, \ldots\ldots, 0.99$, we get the ninety-nine **percentiles.**

Like the median, the p-quantile for a set of observations may not be unique. Now let's see how to compute this p-quantile. When a frequency table for a continuous variable is given, we first decide which of the class-intervals contains z_p. If x_l and x_u are its lower and upper boundaries and F_l and F_u the corresponding cumulative frequencies, then the p-quantile z_p may be determined approximately by using the formula

$$\frac{z_p - x_l}{x_u - x_l} = \frac{np - F_l}{F_u - F_l}$$

or,
$$z_p = x_l + \frac{np - F_l}{f_0} \times c \,, \qquad \ldots \quad (13)$$

where c is the width of the interval and f_0 is its frequency. In (13), if we put $p = \frac{1}{2}$, we get the formula for the median. Now let's use Formula (13) to compute the quartiles in an example.

Example 3 : For the frequency distribution of petiole length of 198 leaves of a pipal tree, the median (i.e., the second quartile) was evaluated in Example 5 of Unit 2. We'll now find the first and third quartiles.

Here $\frac{n}{4} = 49.5$ and $\frac{3n}{4} = 148.5$.

On going through the cumulative frequency table (Table 8, Unit 1), we find that q_1 lies in the interval 3.75 – 4.55 (cm) and q_3 in the interval 6.15–6.95 (cm). So, after putting the appropriate values from Table 7 and 8 of Unit 1 in (13), we get

$$q_1 = 3.75 + \frac{49.5 - 26}{24} \times 0.8$$
$$= 4.533 \text{ cm.}$$

$$q_3 = 6.15 + \frac{148.5 - 145}{33} \times 0.8$$
$$= 6.158 \text{ cm.}$$

See if you can solve this exercise now.

E4) Find the three quartiles for the age-distribution of the Indian population according to the 1981 census (given in Unit 2). Check that $q_2 - q_1 < q_3 - q_2$.

If we are given the first few moments or a small set of quantiles of a frequency distribution, we can get a fairly good idea about the distribution. In fact, for most purposes, it will be enough to state the values of $\bar{x}$, m_2, m_3 and m_4 or those of the three quartiles (or of the nine deciles).

We'll come back to this later. Now we introduce a very useful concept, that of weighted mean.

Weighted Mean

Consider this situation. Students are admitted to a B.Sc. course in Statistics on the basis of their performance in the Higher Secondary, or an equivalent examination. Then don't you think that their scores on the mathematics papers should be considered more important than those on the physics papers? Similarly, shouldn't the scores on language papers be considered least important? It is necessary in such a situation to take into account the relative importance (or weight) of the different observations while evaluating the mean.

Suppose $w_i \geq 0$ is the weight attached to the value x_i (i.e., to the value of x for the ith individual). Then the appropriate mean would be

$$\bar{x}_w = \frac{\sum_1^n w_i x_i}{\sum_1^n w_i} \quad \ldots (15)$$

The measure is called a **weighted mean** of x. This concept is particularly useful in economic studies—in the construction of a price index number. This will become clear from the next example.

Example 4 : The price increases from 1985 to 1989 for five food items have been (in percentage terms) as follows:

132.1 153.4 144.3 119.7 120.1

If the figures given below indicate the relative importance of these items in a typical citizen's diet,

34 19 24 12 11,

then the average price increase for these items should be taken to be

$$\bar{x}_w = \frac{1}{100}\{34 \times 132.1 + 19 \times 153.4 + \ldots\ldots + 11 \times 120.1\}$$

$$= \frac{13626.7}{100} = 136.27 \text{ per cent.}$$

Observe that the formula

$$\bar{x} = \frac{1}{n}\sum_1^k f_i x_i$$

that we used to compute the mean of x from grouped data is nothing but the weighted mean of $x_1, \ldots\ldots, x_k$, the respective frequencies now serving as the weights.

Try this exercise now.

E5) A student gets 85, 76 and 82 marks in the three tests for the course MTE-11. She gets 79 marks in the final examination. What are her average marks if the weightage given to the tests and the final examination are 10, 10, 10 and 70, respectively?

In the next two sections, we'll see how moments and quantiles lead to measures of skewness and kurtosis of a frequency distribution.

3.3 SKEWNESS

In Unit 1, you saw that frequency distributions may be classified as symmetrical and skewed (or asymmetrical). Skewed distributions can again be classified as positively skewed or negatively skewed, according as the longer tail of the distribution is towards the higher or the lower values of the variable (see Fig. 1).

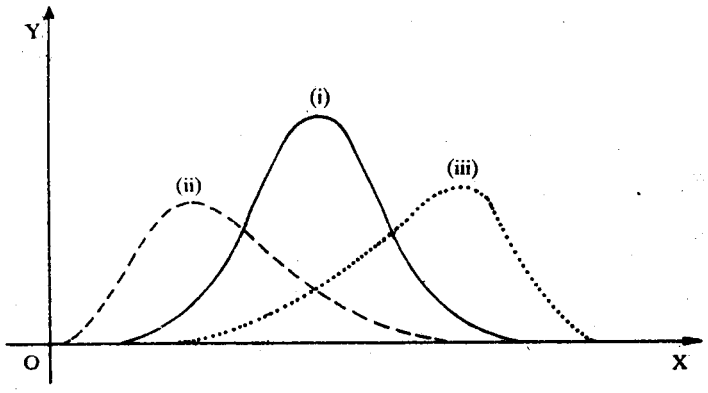


Fig. 1 : (i) Symmetrical (ii) positively skewed and (iii) negatively skewed distributions.

Now, the degree of skewness is the extent to which the given distribution departs from symmetry. A good measure of the degree of skewness has to fulfil the following criteria:

i) It should be a pure number, i.e., should be free of the units in which the variable is measured.

ii) It should be zero, positive and negative for a symmetrical distribution, a positively skew distribution and a negatively skew distribution, respectively.

iii) It should vary between two definite limits, say, $-k$ and $+k$, as the nature of a distribution changes from extreme negative asymmetry to extreme positive asymmetry. Here are some commonly used measures (assuming $s > 0$) :

$$Sk_1 = \frac{(\bar{x} - \overset{\circ}{x})}{s}$$

$$Sk_2 = \frac{3(\bar{x} - \tilde{x})}{s},$$

$$Sk_3 = \frac{(q_3 - q_2) - (q_2 - q_1)}{(q_3 - q_2) + (q_2 - q_1)}$$

and $$Sk_4 = \frac{m_3}{s^3}$$

As usual,
$\bar{x}$ denotes the mean,
$\tilde{x}$ denotes the median,
$\overset{\circ}{x}$ denotes the mode,
q_i, the i^{th} quartile,
m_3, the third moment,
about $\bar{x}$ and s, the standard deviation.

Sk_1 may not be defined, since the mode may not be defined. To get over this difficulty, we use the empirical relation $\bar{x} - \overset{\circ}{x} \simeq 3(\bar{x} - \tilde{x})$ to get the measure Sk_2. Sk_2 and Sk_3 too, may not be unique since the median and the first and third quartiles may not be unique. The square of Sk_4, $Sk_4^2 = \frac{M_3^2}{M_2^3}$ is called Pearson's coefficient and is denoted by b_1.

All the four measures above are free of units. Secondly, for a symmetrical (unimodal) distribution,

$$\text{mean} = \text{median} - \text{mode}$$
and $$q_3 - q_2 = q_2 - q_1 .$$

Further, for most of the distributions which we see in practice, we have the following two observations:

1) For a positively skew distribution,

$$\text{mean} > \text{median} > \text{mode},$$
and $$q_3 - q_2 > q_2 - q_1 .$$

2) For a negatively skew distribution,

$$\text{mean} < \text{median} < \text{mode},$$
and $q_3 - q_2 < q_2 - q_1$. See Fig. 2(a) and (b).

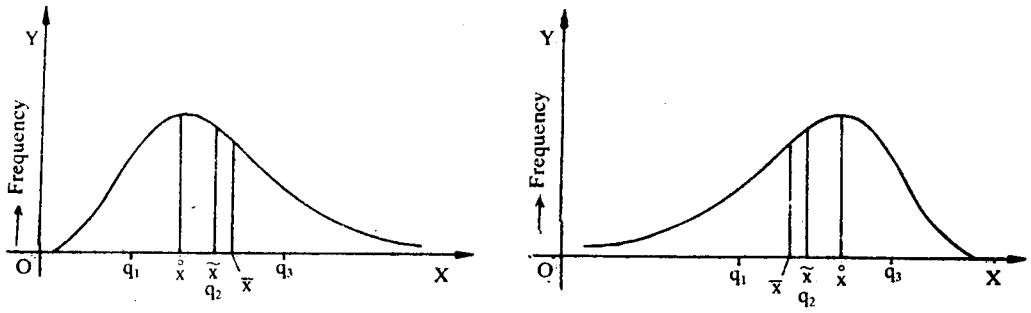


Fig. 2

For a symmetrical distribution, m_3 (or for that matter, any odd-order central moment) is zero. $m_3 > 0$ for a positively skew distribution, and $m_3 < 0$ for a negatively skew distribution. Hence, all the four measures meet the second criterion.

As to the third criterion, we have the following results :

i) For any distribution with $s>0$,

$$-3 \leq Sk_2 \leq +3.$$

ii) For any distribution,

$$-1 \leq Sk_3 \leq +1.$$

Thus, Sk_2 and Sk_3 meet the third criterion. Since we have the empirical relation.

$$\text{mean} - \text{mode} = 3(\text{mean} - \text{median}),$$

we can say that Sk_1, too, roughly meets this criterion. However, $Sk_4 = \sqrt{b_1}$ may take any value between $-\infty$ and $+\infty$ and hence, is inferior to the other measures.

Let's now calculate these four measures for the data on petiole length.

Example 5 : For the frequency distribution of petiole length, we have

$\bar{x} = 5.281$ cm, $\tilde{x}\ (=q_2) = 5.359$ cm, $\mathring{x} = 5.607$ cm.

Also, for this distribution, the first and third quartiles are

$$q_1 = 5.093 \text{ cm}, \quad q_3 = 6.235 \text{ cm},$$

while the standard deviation and third central moment are

$$s = 1.456 \text{ cm}, \quad m_3 = -1.761 \text{ (cm)}^3.$$

Hence,

$$Sk_1 = \frac{(5.281 - 5.607)}{1.456} = -0.224$$

$$Sk_2 = \frac{3 \times (5.281 - 5.359)}{1.456} = -0.461,$$

$$Sk_3 = \frac{(6.235 - 5.359) - (5.359 - 4.533)}{6.235 - 4.553}$$

$$= \frac{-0.05}{1.702} = -0.029, \text{ and}$$

$$Sk_4 = \frac{-1.761}{(1.456)^3} = \frac{-1.761}{1.945} = -0.905.$$

All these values indicate that the distribution is only slightly asymmetric, and that it is a case of negative asymmetry. This is also apparent from the histogram of the distribution (see Fig. 4 in Unit 1).

Now here is an exercise for you.

E6) Show that, for a distribution symmetrical about a, the mean as well as the median is a and the central moments are all equal to zero.

Hint : You may take the values of x to be, say $a \pm h_1$, $a \pm h_2$, ..., $a \pm h_k$ ($h_i \geq 0$ for each i) with frequencies f_i for $a - h_i$ and also for $a + h_i$.

Now that we have seen how to measure the skewness of a frequency distribution, let us talk about its kurtosis.

3.4 KURTOSIS

We now focus attention on another feature of a frequency distribution that determines the shape of the distribution. It is the degree of steepness or pointedness of distribution—or, to use a Greek work, the **kurtosis** of the distribution. Some distributions are flat-topped; some are highly peaked; most distributions will be in between these two extreme types, not too peaked and not too flat-topped either.

In Fig. 3, we have the frequency curves of a distribution that is highly peaked, one that is of moderate kurtosis and a third one which is rather flat-topped.

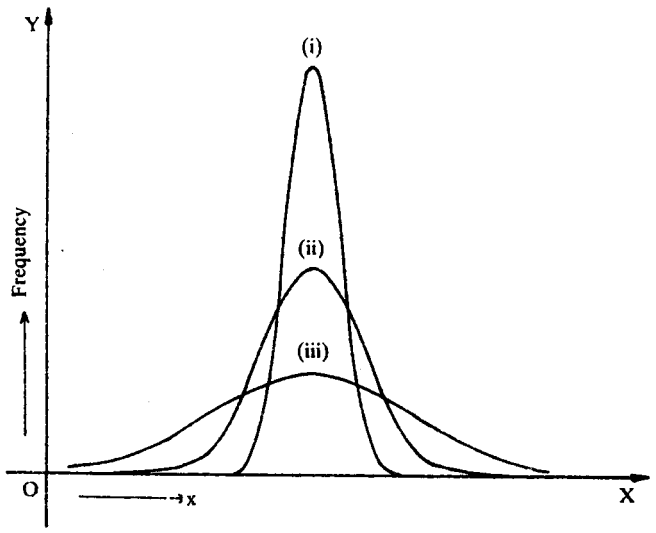


Fig. 3 : Three symmetrical distributions with same mean and s.d. but of varying kurtosis.

It has been observed that for two distributions having the same dispersion and the same degree of skewness, the one with higher kurtosis, usually has higher fourth powers of deviations from the mean and hence a higher value of m_4. This observation is used to define a measure of kurtosis (under the assumption that $s > 0$) as follows :

$$b_2 = m_4/s^4 \qquad \ldots (14)$$

The division by s^4 makes the measure free of units. It also ensures that the measure takes into account that part of the peakedness of the distribution which is independent of (or is in addition to) the part that is due to the variance.

meso means moderate, **lepto** means thin and **platy** mean flat.

For a normal distribution (about which you will learn in Block 3), $b_2 = 3$. This value is taken as a standard against which the kurtosis of other distributions is judged. Any distribution with $b_2 = 3$ is called **mesokurtic** (i.e., of moderate kutosis); one with $b_2 > 3$ is said to be **leptokurtic**, while one with $b_2 < 3$ is said to be **platykurtic**. Thus, in Fig. 3, (i) is lepokurtic, (ii) is mesokurtic, while (iii) is platykurtic.

For any univariate distribution with $s > 0$, we have

$$b_2 \geq 1.$$

Let's prove this.

Proof : Let $u_i = \dfrac{(x_i - \bar{x})}{s}$ for each i.

Then $\sum_1^n u_i = \dfrac{1}{s}\sum_1^n (x_i - \bar{x}) = 0$

and $\sum_1^n u_i^2 = \frac{1}{s^2} \sum_1^n (x_i - \bar{x})^2$

$= \frac{1}{s^2} \times ns^2$

$= n$.

Also, we must have

$\sum_1^n (u_i^2 - 1)^2 \geq 0$, since the L.H.S. is a sum of squares.

This means, $\sum_1^n u_i^4 - 2 \sum_1^n u_i^2 + n \geq 0$.

or $\frac{1}{s^4} \times nm^4 - 2n + n \geq 0$,

or $n(b_2 - 1) \geq 0$.

Since $n > 0$, this implies that $b_2 - 1 \geq 0$, so that the result is established.

Note that we have $b_2 = 1$ iff $n(b_2 - 1) = 0$.

So we can also say that $b_2 = 1$ iff $\sum_1^n (u_i^2 - 1)^2 = 0$, i.e.,

iff $u_i^2 = 1$ for each i,

i.e., iff $x_i = \bar{x} \pm s$ for each i.

Thus, the coefficient of kurtosis $b_2 = 1$, iff the variable x can assume just two values with equal frequencies (so that the mean may be exactly midway between the two values).

Now let's calculate the coefficient of kurtosis for our favourite distribution.

Example 6 : For the frequency distribution of petiole length, we have, from Example 2,

$m_2 = 1.7892$ (cm)^2, $m_4 = 15.535$ (cm)^4.

Hence, the kurtosis coefficient b_2 for this distribution is given by

$b_2 = 15.535/(1.7892)^2$

$= 4.853$.

So, we find that this distribution is slightly leptokurtic.
Try to do this exercise now.

E7) Comment on the skewness and kurtosis of the age-distribution of the Indian population (1981 census) given in E3) in Unit 2, by evaluating appropriate coefficients and also by considering the histogram of the distribution.

Let us now summarise the points covered in this unit.

3.5 SUMMARY

In this unit, you have seen

1) what is meant by moments and quantiles of different orders about A:

$m_r' = \frac{1}{n} \sum_{i=1}^k f_i (x_i - A)^r$ for data in the form of a frequency distribution

and $m_r' = \frac{1}{n} \sum_1^n (x_i - A)^r$ for raw data.

$z_p = x_i + \frac{np - F_1}{f_0} \times c$

2) how central moments, i.e., moments about x are related to moments about any A:

$$m_r = m'_r - \binom{r}{1} m'_{r-1} m'_1 + \binom{r}{2} m'_{r-2} (m'_1)^2 - \ldots\ldots + (-1)^r \binom{r}{r} (m'_1)^r$$

3) when weighting of the observations would be appropriate for the computation of mean,

4) what is meant by skewness and kurtosis :
skewness is the departure from symmetry and kurtosis gives the degree of flatness of a frequency distribution.

5) how to compute measures of skewness and kurtosis of a frequency distribution :

Measures of Skewness :

$$Sk_1 = \frac{(\bar{x} - \overset{\circ}{x})}{s}, \qquad Sk_2 = \frac{3(\bar{x} - \tilde{x})}{s},$$

$$Sk_3 = \frac{(q_3 - q_2) - (q_2 - q_1)}{(q_3 - q_2) + (q_2 - q_1)}, \qquad Sk_4 = \frac{m_3}{s^3}, \text{ provided } s \neq 0.$$

Measure of kurtosis :

$$b_2 = \frac{m_4}{s^4}, (s \neq 0).$$

3.6 SOLUTIONS AND ANSWERS

E1) $\nu'_1 = \frac{1}{n}\sum_1^n u_i$

$$= \frac{1}{n}\sum \frac{(x_i - A)}{c}$$

$$= \frac{1}{cn}\sum (x_i - A)$$

$$= \frac{1}{c}(\bar{x} - A).$$

$\therefore \bar{x} = A + c\nu'_1$

$$\nu'_2 - \nu'_1 = \frac{1}{n}\sum u_i^2 - \left(\frac{1}{n}\sum u_i\right)^2$$

$$= \frac{1}{c^2 n}\sum (x_i - A)^2 - \left[\frac{1}{cn}\sum (x_i - A)\right]^2$$

$$= \frac{1}{c^2}\left[\frac{1}{n}\sum (x_i - A)^2 - (\bar{x} - A)^2\right]$$

$$= \frac{1}{c^2}\left[\frac{1}{n}\sum x_i^2 - \bar{x}^2\right]$$

$$= \frac{1}{c^2} m_2 .$$

Similarly, solve for m_3 and m_4.

E2) $x_i - A = (x_i - \bar{x}) + (\bar{x} - A)$

$$= (x_i - \bar{x}) + d, \text{ say}$$

$$\therefore (x_1 - A)^r = (x_i - \bar{x})^r + \binom{r}{1}(x_i - \bar{x})^{r-1} d + \binom{r}{2}(x_i - \bar{x})^{r-2} d^2$$

$$+ \ldots\ldots + \binom{r}{r-1}(x_i - \bar{x}) d^{r-1} + \binom{r}{r} d^r .$$

On summing over i and dividing the result by n, we get

$$m'_r = m_r + \binom{r}{1} m_{r-1} d + \binom{r}{2} m_{r-2} d^2 + \ldots\ldots + \binom{r}{r-2} m_2 d^{r-2} + \binom{r}{r} d^r.$$

where $d = \bar{x} - A$.

E3) $$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 .$$

Suppose there are k sets of data, each having n_j observations and mean $\bar{x}_j$, j=1, 2,....., k.

Then

$$m_3 = \frac{1}{n}\sum_{j=1}^{k}\left[\sum_{i=1}^{n_j}(x_i - \bar{x}_j + \bar{x}_j - \bar{x})^3\right]$$

$$= \frac{1}{n}\sum_{j}\left[\sum_{i}(x_i - \bar{x}_j)^3 + 3\sum_{i}(x_i - \bar{x}_j)^2(\bar{x}_j - \bar{x}) + 3\sum_{i}(x_i - \bar{x}_j)(\bar{x}_j - \bar{x})^2 + \sum_{i}(\bar{x}_j - \bar{x})^3\right]$$

$$= \frac{1}{n}\sum_{j} n_j\left\{m_{3j} + 3m_{2j}(\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^3\right\},$$

since $\sum_{i}(x_i - \bar{x}_j) = 0$.

Similarly,

$$m_4 = \frac{1}{n}\sum_{j} n_j\left\{m_{4j} + 4m_{3j}(\bar{x}_j - \bar{x}) + 6m_{2j}(\bar{x}_j - \bar{x})^2 + (\bar{x}_j - \bar{x})^4\right\}.$$

E4) $q_1 = 9.41$ years, $q_2 = 20.48$ years, $q_3 = 37.73$ years.

E5) $$\text{Weighted mean} = \frac{10 \times 85 + 10 \times 76 + 10 \times 82 + 70 \times 79}{100}$$

$$= 79.6$$

E6) The first moment about a is

$$m_1' = \frac{1}{n}\left[\sum_{1}^{k}\left\{(a + hi) - a\right\}f_i + \sum_{1}^{k}\left\{(a - hi) - a\right\}f_i\right]$$

$$= \frac{1}{n}\left[\sum_{1}^{k}(hi - hi)\,f_i\right] = 0 .$$

$\Rightarrow \bar{x} = a + m_1' = a.$

If a is not a possible value of x, then the total frequency of values less than a equals the total frequency of values exceeding a. Hence a is the median. If a is a possible value (occurring with frequency f_0, say), then also, total frequency below a equals total frequency above a, hence a is again the median.

E7) The moments about 27.5 (years) are

$m_1' = -0.488 \times 5$

$m_2' = -14.4954 \times 5^2$,

$m_3' = 19.0042 \times 5^3$,

$m_4' = 464.9984 \times 5^4$.

Hence $\bar{x} = 25.06$, $m_2 = 14.2573 \times 5^2$, $m_3 = 39.9931 \times 5^3$, $m_4 = 522.6385 \times 5^4$

$\Rightarrow \sqrt{b_1} = 0.743$, $b_2 = 2.571$.

The distribution is moderately skew and platykurtic.

This is also indicated by the histogram.

UNIT 4 CORRELATION AND REGRESSION

Structure

4.1 INTRODUCTION

Suppose you are given the data on the scores obtained by 30 students in English. From the three units that you have studied so far, you know how to form the frequency distribution of the data. You have also seen how to compute the average score and the standard deviation. Now suppose the scores obtained by these students in Mathematics are also given to you. Apart from finding the average score in Mathematics, you may also like to know whether there is any relationship between a student's score in English and his/her score in Mathematics. It is quite reasonable to think that a student good at English would also be good at Mathematics. But is this fact borne out by the data? In this unit, we are going to discuss some methods which will enable us to answer this question.

We'll start by discussing the tabular and diagrammatic representation of bivariate data, i.e., of data pertaining to two variables. After this, we'll talk about the nature and the degree of relationship between the observations on two variables. We would also like to see if we can build up an equation on the basis of the given data, which can help us in predicting one of the variables when the other is given.

Here are the objectives which you should achieve by the end of this unit.

Objectives

After reading this unit, you should be able to :

- draw a scatter diagram corresponding to the given bivariate frequency distribution
- explain the meaning of "correlation" and "regression"
- fit a regression line to the given data
- compute the correlation coefficient for grouped and ungrouped data
- derive some relationship between the correlation and regression coefficient.

4.2 TABULAR AND DIAGRAMMATIC REPRESENTATION OF BIVARIATE DATA

In this section, we'll briefly discuss how to organise bivariate data. Since we have already discussed the organisation of univariate data in detail (in Unit 1), here we'll only indicate in what way the organisation of bivariate data is different from that of univariate data.

When the number of individuals is not too large, we can present the bivariate data simply in the form of a table with two columns. The values of the two variables, say x and y, for each individual are written down side by side. We may also add a third column, before the other two, to indicate which pair of values relate to which individual. For example, we give the data on the height and yield of dry bark for 18 cinchona plants in this manner in Table 1.

Table 1 : Data on the height and yield of dry bark for 18 cinchona plants

Plant No.	Height of dry bark (inches)	Yield of dry bark (ounces)
1	8	19
2	15	51
3	11	30
4	21	42
5	7	25
6	5	18
7	10	44
8	13	56
9	12	38
10	13	32
11	5	25
12	6	10
13	4	20
14	8	27
15	7	13
16	12	49
17	6	27
18	16	55

Source : *Fundamentals of Statistics* by Goon, Gupta and Dasgupta.

However, as in the univariate case, when the number of individuals is large, the data needs to be grouped into a frequency table. Now in the present case, the frequency table has to be a two-way table, with a suitable number of classes for x and a suitable number of classes for y. In Table 2, you can see that

We decide upon these "suitable" numbers after bearing in mind all the points listed in Sec. 1.3.

- we have divided the number of grains per earhead into 10 classes, 8-12, 13-17, ..., 53-57.
- we have divided the length of earhead into 9 classes, 5.25-6.25, 6.25-7.25, ..., 13.25-14.25.

If there are l classes for x and k classes for y, the table will have kl cells in all. So in Table 2 we have 90 cells.

After specifying the classes we determine the cell frequencies and write them as in Table 2.

Table 2 : Bivariate frequency table for length of earhead and number of grains per ear for 400 ears of a variety of wheat.

	Length (in cm)									
Number of grains	5.25 – 6.25	6.25 – 7.25	7.25 – 8.25	8.25 – 9.25	9.25 – 10.25	10.25 – 11.25	11.25 – 12.25	12.25 – 13.25	13.25 – 14.25	
8-12	1									1
13-17	3	10	4							17
18-22		4	10	9	2					25
23-27			4	41	35	6				86
28-32			1	15	65	37	6	1		125
33-37				9	15	13	21	2		77
38-42					6	17	23	6	3	55
43-47						2	4	3		9
48-52							1	1	2	4
53-57									1	1
Total	4	14	19	74	123	92	55	13	6	400

Source : *Statistical Methods for Agricultural Research Workers*, by Panse and Sukhatme.

Table 2 shows a **bivariate frequency distribution** or a **joint frequency distribution** (say, of x and y). But the row totals or the column totals themselves represent a univariate frequency distribution. If the columns are for x-classes and the rows for y-classes, then

- the column totals give simply the frequency distribution of x. Thus, in Table 2, the last row gives the frequency distribution of the data on length of earhead, divided into the classes, 5.25-6.25, 6.25-7.25, ..., 13.25-14.25.
- the row totals give simply the frequency distribution of y. Again, in Table 2, the last row gives the frequency distribution of the data on number of grains per earhead, divided into the classes, 8-12, 13-17, ..., 53-57.

These are called the **marginal frequency distribution of x** and the **marginal frequency distribution of y.**

But each column or each row of frequencies in the two way table also represents a frequency distribution. Here, in any column, the x value is either fixed or confined to a given interval, but y is allowed to vary. The column of frequencies indicates how the column total (i.e., the frequency for the given x-class) is distributed over the different y-classes. We call this the **conditional** (or array) **frequency distribution** of y, given x. Similarly, each row of frequencies shows the conditional frequency distribution of x, given y.

We could also represent the joint frequency distribution in terms of relative frequencies (obtained by dividing each cell frequency by the total frequency) or cumulative frequencies or cumulative relative frequencies. The cumulative frequencies of the less than type or those of the more than type will be obtained by taking cumulative totals of the frequencies. But here each such total is a double sum. Thus, if f_{ij} is the frequency for the (i,j)th cell, i.e., the cell formed by the ith class of y and the jth class of x, then the cumulative frequency of the less than type for the (i,j)th cell will be

$$F_{ij} = \sum_{i' \leq i} \sum_{j' \leq j} f_{i'j'} \quad (i=1,2,\ldots., k; j=1,2,\ldots\ldots,l)$$

and the cumulative frequency of the more-than type for the (i, j)th cell will be

$$F'_{ij} = \sum_{i' \geq i} \sum_{j' \geq j} f_{i'j'} \quad (i=1,2,\ldots\ldots k; j=1,2,\ldots\ldots,l).$$

Let us now look at the ways in which a bivariate distribution can be represented diagrammatically.

Diagrammatic Representation

Like a univariate frequency distribution, a bivariate frequency distribution too may be represented diagrammatically.

If raw data are to be represented, then we make use of what is called a **scatter diagram** (or **dot diagram**). Here, the n pairs of values, say (x_i, y_i) for i=1, 2, ..., n, of the variables x and y are plotted as points w.r.t. a rectangular system of coordinates. In Fig. 1, we have the scatter diagram for the data in Table 1.

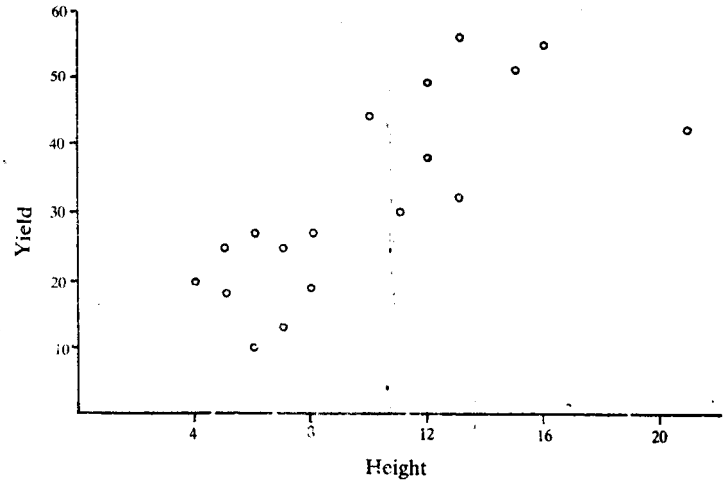


Fig. 1 : Scatter diagram for the data in Table 1.

When the data are in grouped form, we use a three-dimensional analogue of a histogram (in case both the variables are continuous) or a three-dimensional analogue of a column diagram or a frequency polygon (in case the variables are both discrete). In the first case we take two mutually perpendicular axes of coordinates on a plane, one for x and the other for y. Then we can show the class-intervals on each axis. This gives us a network of rectangles, each corresponding to a cell of the frequency-table. Next, we take an axis perpendicular to the xy-plane to represent frequency density. Finally, on each rectangle as base, we erect a block, i.e., a parallelopiped (see Fig. 2). The height of this box is equal to the frequency density.

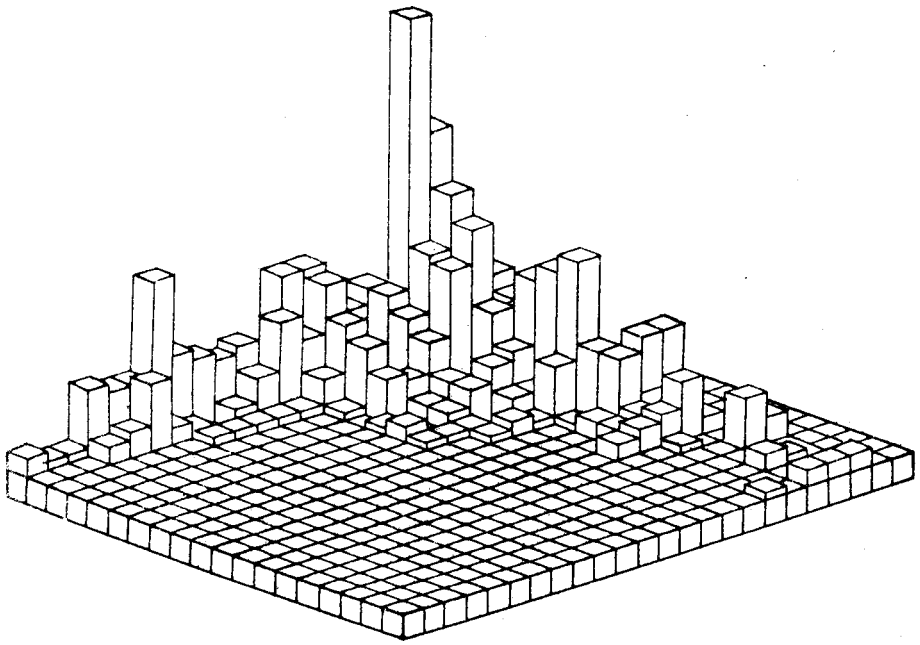


Fig. 2

Note that in the bivariate case,

$$\text{frequency density for a cell} = \frac{\text{cell frequency}}{\text{area of cell}}$$

Even when the variable are discrete, we take three axes. We indicate the possible distinct values of x on the x-axis and those of y on the y-axis. At the point (x_i, y_j), we now erect a column of height f_{ij} to get the column diagram as in Fig. 3.

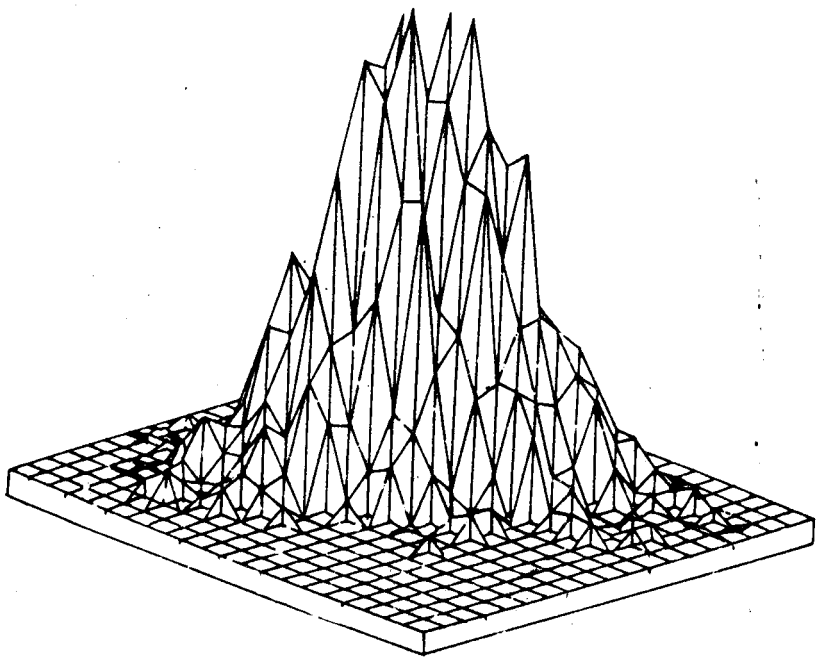


Fig. 3 : Column diagram and frequency polygon

In order to get the frequency polygon, on the other hand, we join the tops of adjacent columns by straight-line segments.

The three-dimensional analogues of a frequency polygon as well as a histogram are called **stereograms.**

In this unit, we'll be using scatter diagrams to represent bivariate data. So, before going any further, see if you can draw a scatter diagram on the basis of given data.

E1) Draw a scatter diagram on the basis of the following data on the size of crop and percentage of wormy fruits during a season for 12 apple trees.

Tree No.	Size of crop (i.e., number of apples in hundreds)	Percentage of wormy fruits
1	8	59
2	6	58
3	11	56
4	22	53
5	14	50
6	17	45
7	18	43
8	24	42
9	19	39
10	23	38
11	26	30
12	40	22

Source : *Statistical Methods,* by Snedecor and Cochran

Now, on the basis of such bivariate data, let us see if we can establish a relationship between the two variables under study.

4.3 WHAT DO WE MEAN BY REGRESSION ANALYSIS?

Consider the following situations :

i) The advertising manager of a firm collects data about the money spent on advertising and the sales in each year during 1980-90.

ii) A doctor collects data about the extent of cellular damage induced by exposure to differing intensities of radiation.

iii) A social worker collects data about the number of children, the ages of parents at the time of marriage and their educational status.

In each of these cases, data are collected to explore the possible relationship between the variables. The advertising manager wants to know whether there is any relationship between the money spent on advertising and the sale figures. He would also like to know what the relationship is, for it will help him decide how much more money he should spend to reach a particular target of sales.

Similarly, the doctor is concerned about the extent of damage caused by exposure to radiation, and would want to have a clear idea before prescribing the dose.

The social worker wants to know what kind of relationship, if any, exists between the number of children born to a couple, and the ages of the parents at the time of their marriage and also their educational status.

Under **regression analysis,** we deal with statistical mehods which help us in formulating models which describe relationships among variables. These models are eventually used for prediction. The term **simple regression** is used when we are exploring the relationship between two variables (as in the first two situations above). When we are predicting one variable on the basis of information on more than one 'predictor' variables (as in the third situation), we use the term, **multiple regression.** In this course, we'll only talk about simple regression.

Now suppose we have bivariate data and want to investigate the relationship between the two variables. The first step is to draw a scatter diagram. By looking at it, we can get some idea about the relationship. Here are some scatter diagrams pertaining to different sets of data (Fig. 4(a) – (f)). We'll now briefly explain how to interpret these scatter diagrams.

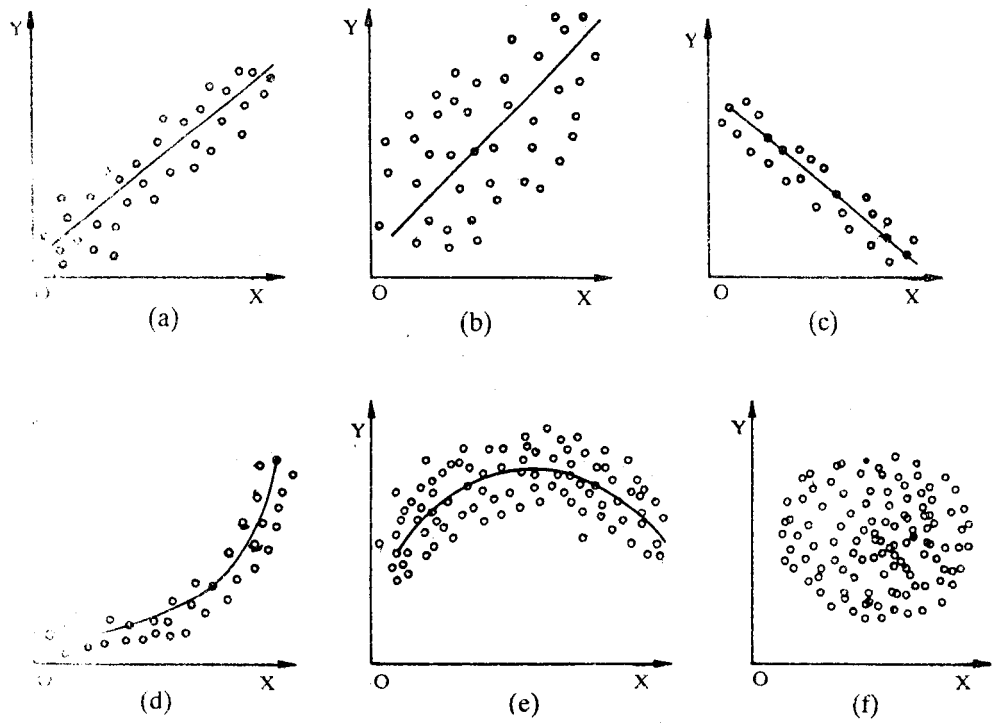


Fig. 4

In (a), (b) and (c), you can see that all the points are quite close to a straight line path. In (a) and (b), this straight line has positive slope, while the one in (c) has negative slope. We say that the data in (a) and (b) show a **positive linear relationship.** Of course, the points in (b) are more scattered than those in (a). The scatter diagram in (c) exhibits a **negative linear relationship.** From (d) and (e), we can say that the data indicate curvilinear relationships. But, from the scatter diagram in (f), we cannot think of any relationship existing between the variables in the data.

For bivariate data, we know that the conditional distribution of y given x indicates how the frequency for a given x is distributed over different classes of y. Now, if the conditional distribution of y given x does not change significantly for different values of x, then we can say that there is no relationship between y and x. And, in this case, the x-values are useless for predicting y. One way to see if there is some kind of relationship between y and x is to consider the average value of y given the values of x. In other words, we consider the conditional average of y given x. If we plot the conditional average of y given x, against x, we get a curve which is called the **regression curve** of y on x. In general, this curve could be quite complicated. So, our first reaction would be to explore the possibility of approximating it by a simple curve like a straight line.

There is yet another motivation for using the regression curve as defined above. Recall that in Unit 2, we have seen that

$$\sum_{i=1}^{n} (x_i - A)^2 \text{ is least when } A = \bar{x}.$$

So,

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - A)^2 \text{ is least when } A = \bar{x}. \qquad \ldots (1)$$

Now, let us again refer to the data on length of earhead (x) and number of grains per ear (y) given in Table 2. Suppose we are interested in predicting the number of grains per ear on the basis of ear length. Suppose further, that the loss we suffer due to wrong prediction is proportional to the squared error, where error is the difference between observed and predicted values. Then our average loss will be minimum, if the average squared error is minimum. Using (1), we can say that the average squared error will be least if we take the mean number of grains per ear for any specific length

Read the adjoining paragraph slowly and carefully to understand the argument.

of the ear, as our predicted value. We can read out this predicted value from the regression curve.

The term 'regression' was first used by Sir Francis Galton in 1877, in connection with his study of human height. He found that the height of the children of tall parents tended to regress (or move back) towards the average height of the population. Galton called the line describing this relationship, a 'line of regression', and the terminology has since been accepted universally.

In the next section, we'll see how to fit a regression line to given data.

4.4 SIMPLE REGRESSION LINE

In this section, we describe a method of fitting a straight line equation to the given bivariate data. This method will lead to the actual regression line if the regression is, indeed, linear. Otherwise, it leads to the best linear approximation to the true regression curve.

Now, in the light of what we have said towards the end of the last section, we try to fit such a line that the average squared error is minimum. This indicates that the straight line be fitted by the **method of least squares**. That is, we choose the constants a and b in the regression equation $y = a + bx$ in such a way that

$$\sum (y_i - a - bx_i)^2 \text{ is a minimum.}$$

Let e_i = observed value of y – predicted value of y (corresponding to x_i).

$= y_i - (a + bx_i)$, $i=1, 2, \ldots\ldots, n.$

e_i is actually the error of prediction for the pair (x_i, y_i).

For obtaining a and b, we minimise $\sum e_i^2$. Recall (MTE-07, Unit 8) that a necessary condition for minimum is that the first partial derivatives should vanish. Thus, equating to zero, the partial derivatives of $\sum e_i^2$ w.r.t. a and b, we get

$$-2\sum e_i = 0 \text{ and } -2\sum e_i x_i = 0.$$

$$\Rightarrow \sum e_i = \sum (y_i - a - bx_i) = 0 \qquad \ldots (2)$$

$$\text{and } \sum e_i x_i = \sum x_i (y_i - a - bx_i) = 0. \qquad \ldots (3)$$

If, by $\hat{a}$ and $\hat{b}$, we denote the solutions of Equations (2) and (3), we get

$$\sum y_i - n\hat{a} - \hat{b}\sum x_i = 0, \text{ from (2).}$$

$$\therefore \bar{y} - \hat{a} - \hat{b}\bar{x} = 0,$$

$$\text{or } \hat{a} = \bar{y} - \hat{b}\bar{x}. \qquad \ldots (4)$$

Similarly, from (3), we get

$$\sum x_i y_i - \hat{a}\, n\bar{x} - \hat{b}\sum x_i^2 = 0.$$

Substituting the value of $\hat{a}$ from (4), we get

$$\sum x_i y_i - (\bar{y} - \hat{b}\bar{x})\, n\bar{x} - \hat{b}\sum x_i^2 = 0.$$

$$\Rightarrow \hat{b}\left[\sum x_i^2 - n\bar{x}^2\right] = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\text{or, } \hat{b}\sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x})(y_i - \bar{y}) \qquad \ldots (5)$$

We have already proved the equality of $\sum x_i^2 - n\bar{x}^2$ and $\sum (x_i - \bar{x})^2$ in Unit 2. On exactly similar lines, you can show the equivalence of $\sum (x_i - \bar{x})(y_i - \bar{y})$ and $\sum x_i y_i - n\bar{x}\bar{y}$. We leave this as an exercise to you (see E2). (5) gives us,

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}, \qquad \ldots (6)$$

where $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$ and $S_x^2 = \sum (x_i - \bar{x})^2$.

E2) Prove that $\sum_{1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$

To see that the solutions (4) and (6), indeed, minimise $\sum e_i^2$, we proceed as below :

We have

$$e_i = y_i - a - bx_i$$
$$= (y_i - \bar{y}) - b(x_i - \bar{x}) + (\bar{y} - a - b\bar{x})$$

Therefore,

$$e_i^2 = (y_i - \bar{y})^2 + b^2(x_i - \bar{x})^2 + (\bar{y} - a - b\bar{x})^2 - 2(y_i - \bar{y})\, b\,(x_i - \bar{x}) +$$
$$2(y_i - \bar{y})(\bar{y} - a - b\bar{x}) - 2b(x_i - \bar{x})(\bar{y} - a - b\bar{x}) \quad \ldots (7)$$

Now, when we sum over i=1, 2, ..., n, the last two terms on the right hand side in (7) vanish, since $\sum_{1}^{n} (y_i - \bar{y}) = 0$ and $\sum_{1}^{n} (x_i - \bar{x}) = 0$

Hence we get,

$$\sum_{1}^{n} e_i^2 = \sum_{1}^{n} (y_i - \bar{y})^2 + b^2 \sum_{1}^{n} (x_i - \bar{x})^2 + \sum_{1}^{n} (\bar{y} - a - b\bar{x})^2 - 2b \sum_{1}^{n} (y_i - \bar{y})(x_i - \bar{x}).$$

Let us write $\sum (y_i - \bar{y})^2$ as S_y^2.

Then

$$\sum_{1}^{n} e_i^2 = n(\bar{y} - a - b\bar{x})^2 + S_y^2 + b^2 S_x^2 - 2bS_{xy}$$

Have you noted that the expressions of $\frac{S_y^2}{n}$ and $\frac{S_x^2}{n}$ are similar to the variances of y and x?

$$= n(\bar{y} - a - b\bar{x})^2 + (b^2 S_x^2 - 2bS_{xy} + \frac{S_{xy}^2}{S_x^2}) + S_y^2 - \frac{S_{xy}^2}{S_x^2}, \text{ provided } S_x \neq 0.$$

$$= n(\bar{y} - a - b\bar{x})^2 + (bS_x - \frac{S_{xy}}{S_x})^2 + (S_y^2 - \frac{S_{xy}^2}{S_x^2}) \quad \ldots (8)$$

The last term in (8) does not depend on a and b. You would agree that the first two terms will have the smallest value (zero) if we set

$$b = \frac{S_{xy}}{S_x^2} = \hat{b} \text{ and } a = \bar{y} - b\bar{x} = \hat{a}.$$

This shows that the solutions (4) and (6) do minimise $\sum e_i^2$.

So, for the given data, we have to first calculate $\bar{x}$, $\bar{y}$, S_x^2 and S_{xy}. Then we get the values of a and b and hence, fix the regression line : $y = a + bx$. Here, x is called the **predictor variable** while y is said to be the **predictant**. The constant a is the y-intercept of the line and represents the predicted value of y when x = 0. The constant b, which is the slope of the fitted line, is the rate of increase of the predicted value of y per unit increase in the value of x. It is called the **regression coefficient of y on x** and is also denoted by $\mathbf{b_{yx}}$. In Galton's example, b_{yx} was negative. This led to the regression phenomenon.

Sometimes x is also called the **independent** variable and y, the **dependent** variable.

Now, we know that $S_x^2 = \sum_{1}^{n} (x_i - \bar{x})^2 = \sum_{1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}$, and

$$S_y^2 = \sum_{1}^{n} y_i^2 - \frac{(\sum_{1}^{n} y_i)^2}{n}$$

Further, in E2) you have proved that

$$S_{xy} = \sum_{1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{1}^{n} x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Putting these expressions for S_x^2 and S_{xy} in the formula for b_{yx}, we get

$$b_{yx} = \frac{\sum_1^n x_i y_i - (\sum_1^n x_i)(\sum_1^n y_i)/n}{\sum_1^n x_i^2 - (\sum_1^n x_i)^2/n}$$

or,

$$b_{yx} = \frac{n\sum_1^n x_i y_i - (\sum_1^n x_i)(\sum_1^n y_i)}{n\sum_1^n x_i^2 - (\sum_1^n x_i)^2} \quad \ldots (9)$$

Suppose we look upon x as the dependent variable. Then to predict the values of x from the values of y, we use the regression equation,

$$x = a' + b'y ,$$

where $b' = \frac{S_{xy}}{S_y^2} (S_y \neq 0)$,

and $a' = \bar{x} - b'\bar{y}$.

b', the regression coefficient of x on y is denoted by b_{xy}, and we have

$$b_{xy} = \frac{n\sum_1^n x_i y_i - (\sum_1^n x_i)(\sum_1^n y_i)}{n\sum_1^n y_i^2 - (\sum_1^n y_i)^2} \quad \ldots (10)$$

Using Formulas (9) and (10), you can calculate b_{yx} and b_{xy} from raw data. But sometimes you may need to calculate b_{xy} or b_{yx} from grouped data. We now give the formulas which you can use in such cases.

Let us first fix a notation. We write

$$f_{i0} = \sum_j f_{ij} \quad (i=1, 2, \ldots\ldots, k), \text{ and}$$

$$f_{0j} = \sum_j f_{ij} \quad (j=1, 2, \ldots\ldots, l).$$

Now the formulas for the regression coefficients of grouped data are ,

$$b_{yx} = \frac{\sum_i \sum_j f_{ij} (x_i - \bar{x})(y_j - \bar{y})}{\sum_i f_{i0} (x_i - \bar{x})^2} ,$$

$$= \frac{\sum_i \sum_j f_{ij} x_i y_j - (\sum_i f_{i0} x_i)(\sum_j f_{0j} y_j)/n}{\sum_i f_{i0} x_i^2 - (\sum_i f_{i0} x_i)^2/n} , \quad \ldots (11)$$

and,

$$b_{xy} = \frac{\sum_i \sum_j f_{ij} x_i y_j - (\sum_i f_{i0} x_i)(\sum_j f_{0j} y_j)/n}{\sum_j f_{0j} y_j^2 - (\sum_j f_{0j} y_j)^2/n} . \quad \ldots (12)$$

Now let us use Formula (9) to fit a regression line to the raw data in the next example

Example 1 : Look at the data given the following table, where x is the independent variable and y is the dependent one.

x_i	1	1	2	3	4	4	5	6	6	7
y_i	2.1	2.5	3.1	3.0	3.8	3.2	4.3	3.9	4.4	4.8

Now, to find the values of a and b in the equation of the line of regression,

$$y = ax + b,$$

we form the following table.

x_i	y_i	x_i^2	x_iy_i
1	2.1	1	2.1
1	2.5	1	2.5
2	3.1	4	6.2
3	3.0	9	9.0
4	3.8	16	15.8
4	3.2	16	12.8
5	4.3	25	21.5
6	3.9	36	23.4
6	4.4	36	26.4
7	4.8	49	33.6
39	35.1	193	152.7

Here, we have $\sum x_iy_i = 152.7$, $\sum x_i = 39$

$\sum y_i = 35.1$, $\sum x_i^2 = 193$ and $n = 10$.

Putting these values in (9), we get

$$b_{yx} = \frac{10 \times 152.7 - 39 \times 35.1}{10 \times 193 - (39)^2}$$

$$= 0.387.$$

Further, $\bar{x} = \frac{39}{10} = 3.9$ and $\bar{y} = 3.51$.

Therefore, $a = \bar{y} - b\bar{x} = 3.51 - (0.387)\ 3.9$

$$= 2.00 .$$

Hence, the regression line is $y = 2 + 0.387x$.

In Fig. 5, you can see the scatter diagram and the regression line for these data.

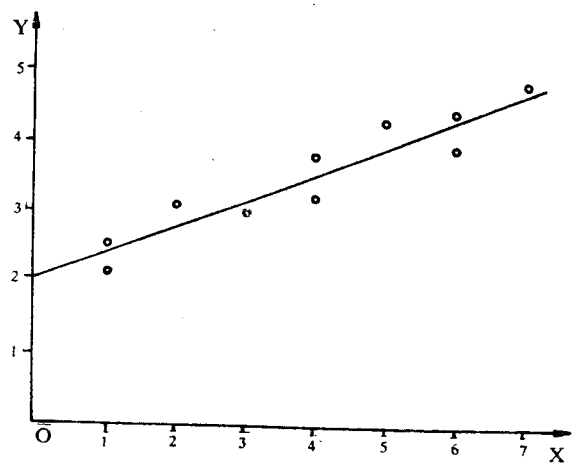


Fig. 5 : Regression line for the data in Example 1

If you have followed the discussion in this section, you should be able to do this exercise.

E3) A preparation of insulin was being studied to determine its effect on reducing the blood-sugar level in rats. Seven rats were injected with different dosages. Reductions in their blood-sugar levels are :

Dosage	.20	.25	.25	.30	.40	.50	.50
Reduction in blood sugar	30	26	40	35	54	56	65

a) Identify the dependent and independent variables.

b) Plot the scatter diagram.

c) Find the equation to the regression line and plot it on the same graph as b).

As usual, you can expect a suitable change of origin and scale to simplify your calculations. In fact, if

$$u = \frac{x-A}{c} \text{ and } v = \frac{y-B}{d}, \text{ then}$$

$\bar{y} = B + d\bar{v}$, $\bar{x} = A + c\bar{u}$ and $b_{yx} = \frac{d}{c} b_{vu}$.

Use this result while solving this exercise now.

E4) Find the regression line, $y = a + bx$, corresponding to the following data :

x	170	147	166	125	182	133	146	125	136	179
y	698	518	725	485	745	538	485	625	471	798

So far you have seen how to find the line of regression to fit the given data. Using this line, we can then predict the value of the dependent variable for given values of the independent variable. For example, the regression line for the data in Example 1 is given by $y = 2 + 0.387x$. We could use this to predict the value of y, when x is, say, 3.2. Thus, we can expect that

$$y = 2 + 0.387 \times 3.2 = 3.24.$$

But we should also know how good our estimate is. A measure of 'goodness' of our prediction is given by the correlation coefficient and this is what we are going to study in the next section.

4.5 CORRELATION COEFFICIENT

In the previous section, we saw how to fit a regression line of y on x, $y = a + bx$ to bivariate data. Recall that we have obtained the "estimates" of a and b as

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{S_{xy}}{S_x^2}.$$

We had also seen that with these values of $\hat{a}$ and $\hat{b}$, we get the least squared error. Let us define

$$\hat{e}_i = y_i - \hat{a} - \hat{b}x_i \text{ for } i=1, 2, \ldots\ldots, n.$$

Now, if you go back to (8), you will see that

$$\sum_{i=1}^{n} \hat{e}_i^2 = S_y^2 - \frac{S_{xy}^2}{S_x^2}, \text{ since the first two terms on the right hand side of (8)}$$

vanish for $\hat{a}$ and $\hat{b}$.

Let us write $r = \frac{S_{xy}}{S_x S_y}$. Then

$$\frac{S_{xy}}{S_x} = r S_y \text{ and we have}$$

$$\sum_{1}^{n} \hat{e}_i^2 = S_y^2 (1-r^2) \qquad \ldots (13)$$

The left hand side of (13) is non-negative. This implies that $1-r^2$ should also be non-negative. Thus,

$$0 \leq r^2 \leq 1.$$

Now, if $r^2 = 1$, $\sum_1^n \hat{e}_i^2$ will be zero. This means the prediction error is zero if $r^2=1$. On the other hand, if r^2 is close to zero, $\sum_1^n \hat{e}_i^2$ will be close to S_y^2. Thus, we can take r^2 (or,r) to be a measure of the "goodness' of the fitted regression line in explaining the true regression of y on x. In fact, r measures the strength of the linear relationship between y and x. The quantity $r = \frac{S_{xy}}{S_x S_y}$ is called the **correlation coefficient** between x and y.

We have noted earlier that $s_x^2 = \frac{S_x^2}{n}$ and $s_y^2 = \frac{S_y^2}{n}$ are the variances of x and y, respectively. We call $s_{xy} = \frac{S_{xy}}{n}$, **the covariance between x and y (Cov(x,y)).** Thus, we can define the correlation coefficient between x and y as

$$r = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}(x) \cdot \text{Var}(y)}}$$

Since $0 \leq r^2 \leq 1$, it follows that $-1 \leq r \leq 1$.

The correlation coefficient, r and the regression coefficient of y on x (or, of x on y), b_{yx} (b_{xy}) have the same sign; that of S_{xy}. So, if an increase in x is associated with an increase in y, we would have a positive covariance and consequently, a positive correlation coefficient. On the other hand, if an increase in x is associated with a decrease in y, the sign of the correlation coefficient would be negative.

Now, $y_i = \hat{a} + \hat{b}x_i + \hat{e}_i$, $i=1, 2, \ldots\ldots, n$.

Let us find the variance of y. For this, we first write

$$y_i - \bar{y} = \hat{b}(x_i - \bar{x}) + \hat{e}_i, \quad \text{since } \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

$$\text{Now, } S_y^2 = \sum_1^n (y_i - \bar{y})^2 = \hat{b}^2 \sum_1^n (x_i - \bar{x})^2 + \sum_1^n \hat{e}_i^2 + 2\hat{b} \sum_1^n (x_i - \bar{x}) \hat{e}_i .$$

$$= \hat{b}^2 S_x^2 + S_y^2 (1-r)^2 + 2\hat{b} \sum (x_i - \bar{x}) \hat{e}_i \quad \text{by (13)}$$

$$\text{But } \sum_1^n (x_i - \bar{x})\hat{e}_i = \sum_1^n (y_i - \bar{y})(x_i - \bar{x}) - \hat{b} \sum_1^n (x_i - \bar{x})^2$$

$$= S_{xy} - \hat{b} S_x^2$$

$$= 0.$$

$$\text{Therefore, } S_y^2 = \hat{b}^2 S_x^2 + S_y^2 (1-r^2)$$

$$= \frac{S_{xy}^2}{S_x^2} + S_y^2 (1-r^2), \quad \text{since } \hat{b} = \frac{S_{xy}}{S_x^2}$$

$$= r^2 S_y^2 + S_y^2 (1-r)^2, \text{ since } r = \frac{S_{xy}}{S_x S_y}$$

$$\text{Thus, } s_y^2 = r^2 s_y^2 + s_y^2 (1-r^2) \quad \ldots (14)$$

We see here that s_y^2, the variance of y is the sum of two parts. The first term on the right hand side of (14) is the variance of $\hat{a} + \hat{b}x_i$, which is the part of y explained by the linear association of y and x. The second term is the variance of the residuals $\hat{e}_i$. This shows that r^2 can be interpreted as the proportion of variability in y, which is explained by its linear association with x.

$$\text{Now, } r = \frac{S_{xy}}{S_x S_y}.$$

An alternative expression is

$$r = \frac{\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)/n}{\sqrt{\sum_i x_i^2 - (\sum_i x_i)^2/n}\ \sqrt{\sum_i y_i^2 - (\sum_i y_i)^2/n}}$$

We get these expressions by using a technique which we had used earlier to get Formulas (9) and (10).

$$= \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\sqrt{n\sum_i x_i^2 - (\sum_i x_i)^2}\ \sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}} \quad \ldots (15)$$

Formula (15) is very useful if we want to compute the correlation coefficient from raw data. Now let us derive a formula for the correlation coefficient for grouped data

If x_i is the class-mark of the ith class of x, and y_j is that of the jth class of y, then we have

$$\text{Var}(x) = \frac{1}{n}\sum_i f_{i0}\,(x_i - \bar{x})^2$$

$$= \frac{1}{n}\left\{\sum_i f_{i0}\,x_i^2 - \frac{1}{n}(\sum_i f_{i0}x_i)^2\right\}$$

Recall that $f_{i0} = \sum_j f_{ij}$ and $f_{0j} = \sum_i f_{ij}$

$$\text{Var}(y) = \frac{1}{n}\sum_j f_{0j}\,(y_j - \bar{y})^2$$

$$= \frac{1}{n}\left\{\sum_j f_{0j}\,y_j^2 - \frac{1}{n}(\sum_j f_{0j}y_j)^2\right\}$$

$$\text{and Cov}(x,y) = \frac{1}{n}\sum_i\sum_j f_{ij}\,(x_i - \bar{x})\,(y_j - \bar{y})$$

$$= \frac{1}{n}\left\{\sum_i\sum_j f_{ij}x_i y_j - \frac{1}{n}\Big(\sum_i f_{i0}x_i\Big)\Big(\sum_j f_{0j}y_j\Big)\right\}$$

Now we are going to introduce some notation which will help us in the calculations later.

Let $\quad X_j = \sum_i f_{ij}\,x_i$

and $\quad Y_i = \sum_j f_{ij}\,y_i$

So that $\quad \sum_j X_j = \sum_j\sum_i f_{ij}x_i = \sum_i x_i \sum_j f_{ij} = \sum_i f_{i0}x_i$,

$$\sum_i Y_i = \sum_j f_{0j}\,y_j$$

and $\quad \sum_j X_j\,y_j = \sum_i x_i Y_i = \sum_i\sum_j f_{ij}x_i y_j$.

So, the formula for the correlation coefficient becomes

$$r = \frac{\sum_i x_i Y_i - (\sum_j X_j)(\sum_i Y_i)/n}{\sqrt{\sum_i f_{i0}x_i^2 - \Big(\sum_j X_j\Big)^2/n}\ \sqrt{\sum_j f_{0j}y_j^2 - \Big(\sum_i Y_i\Big)^2/n}}$$

$$= \frac{n\sum_i x_i Y_i - (\sum_j X_j)(\sum_i Y_i)}{\sqrt{n\sum_i f_{i0}x_i^2 - \left(\sum_j X_j\right)^2}\ \sqrt{n\sum_j f_{0j}y_j^2 - \left(\sum_i Y_i\right)^2}}$$

We'll use this formula to calculate the correlation coefficient in Example 3. But, first we give an example about ungrouped data.

Example 2 : Consider the data of Table 1. Let us denote by x, the height (in inches) and by y, the yield of dry bark (in ounces) per cinchona plant. It would be interesting to know how closely the two are related. The scatter diagram (Fig. 1) indicates moderate positive correlation. To obtain the correlation coefficient, we do the preliminary computations in the table below.

x_i	y_i	x_i^2	y_i^2	x_iy_i
8	19	64	361	152
15	51	225	2601	765
11	30	121	900	330
21	42	441	1764	882
7	25	49	625	175
5	18	25	324	90
10	44	100	1936	440
13	56	169	3136	728
12	38	144	1444	456
13	32	169	1024	416
5	25	25	625	125
6	10	36	100	60
4	20	16	400	80
8	27	64	729	216
7	13	49	169	91
12	49	144	2401	588
6	27	36	729	162
16	55	256	3025	880
179	581	2133	22293	6636

Since

$$n\sum_1^n x_i^2 - (\sum_1^n x_i)^2$$

$$= 18 \times 2133 - (179)^2$$

$$= 38394 - 32041 = 6353,$$

$$n\sum_1^n y_i^2 - (\sum_1^n y_i)^2$$

$$= 18 \times 22293 - (581)^2$$

$$= 401274 - 337561 = 63713$$

and

$$n\sum_1^n x_i y_i - (\sum_1^n x_i)(\sum_1^n y_i)$$

$$= 18 \times 6636 - 179 \times 581$$

$$= 119448 - 103999 = 15449,$$

The correlation coefficient between height and yield of dry bark per plant is

$$r_{xy} = \frac{15449}{\sqrt{6353 \times 63713}} = \frac{15449}{20119} = 0.768$$

Now we take an example of grouped data. To simplify the calculations here, we shall make use of our usual technique: change of origin and scale. So let's first outline the technique here.

If $u = a + bx$ $(b \neq 0)$, $v = c + dy$ $(d \neq 0)$, and the correlation coefficient of x and y is defined and is r_{xy}, then the correlation coefficient of u and v is also defined and is

$$r_{uv} = \pm r_{xy}, \qquad \ldots (16)$$

according as b and d have the same signs or have opposite signs. We are leaving the proof of this statement to you as an exercise. See E5). To solve it, you will have to show that

$$s_u^2 = b^2 s_x^2, \; s_v^2 = d^2 s_y^2 \text{ and } \mathrm{Cov}(u,v) = bd\, \mathrm{Cov}(x,y).$$

Let's look at an example now.

Example 3 : Consider the grouped data of Table 2. Let us find r_{xy}, where x is the length (in cm) and y is the number of grains per earhead of wheat. To compute this coefficient, let us make a change of base and scale for each variable. Precisely, let

$$u = (x - x_0)/b, \; v = (y - y_0)/d, \qquad \ldots (17)$$

where $x_0 = 9.75$, $b = 1$ (class-width for any x class), $y_0 = 30$ and $d = 5$ (class-width for any y class). The computations needed to obtain r_{uv} are shown below (taking x_i to be the class-mark of the ith x-class and y_j to be that of the jth y-class, u_i and v_j being defined by (17)).

u_i \ v_j	−4	−3	−2	−1	0	1	2	3	4	f_{0j}	$f_{0j}v_j$	$f_{0j}v_j^2$	U_j	U_jv_j
−4	1	–	–							1	−4	16	−4	16
−3	3	10	4	–	–					17	−51	153	−50	150
−2	–	4	10	9	2	–				25	−50	100	−41	82
−1		–	4	41	35	6	–	–		86	−86	86	−43	43
0			1	15	65	37	6	1		125	0	0	35	0
1			–	9	15	30	21	2	–	77	77	77	69	69
2				–	6	17	23	6	3	55	110	220	93	186
3					–	2	4	3	–	9	27	81	19	57
4						–	1	1	2	4	16	64	13	52
5							–	–	1	1	5	25	4	20
f_{i0}	4	14	19	74	123	92	55	13	6	400	44	822	95	675
$f_{i0}u_i$	−16	−42	−38	−74	0	92	110	39	24	95				
$f_{i0}u_i^2$	64	126	76	74	0	92	220	117	96	865				
V_i	−13	−38	−36	−50	−12	64	83	27	19	44				
u_iV_i	52	114	72	50	0	64	166	81	76	675				

Recall that $U_j = \sum_i f_{ij}u_i$ and $V_i = \sum_j f_{ij}v_j$.

Now, $\quad n\sum_i f_{i0}u_i^2 - \left(\sum_j U_j\right)^2 = 400 \times 865 - (95)^2$

$$= 346000 - 9025 = 336975.$$

$$n\sum_j f_{0j}v_j^2 - \left(\sum_i V_i\right)^2 = 400 \times 822 - (44)^2$$

$$= 328800 - 1936 = 326864$$

$$\text{and } n\sum_i u_i V_i - \left(\sum_j U_j\right)\left(\sum_i V_i\right) = 400 \times 675 - 95 \times 44$$

$$= 270000 - 4180 = 265820$$

Hence,

$$r_{uv} = \frac{265820}{\sqrt{336975 \times 326864}}$$

$$= \frac{265820}{331881} = 0.801.$$

Since b and d are both of the same sign, the correlation coefficient between x and y, r_{xy}, has also the value 0.801.

Try to do the following exercises now. By doing them, you will gain a better understanding of the concepts covered in this section.

E5) If $u = a + bx\ (b \neq 0)$, $v = c + dy\ (d \neq 0)$ and if the correlation coefficient of x and y exists and is equal to r_{xy}, then the correlation coefficient of u and v also exists and is given by

$$r_{uv} = \pm r_{xy},$$

according as b and d have the same signs or have opposite signs.

E6) The correlation coefficient between x and y is -0.73. What is then the correlation coefficient between

a) $x + 5$ and $y - 4$?

b) $2x - 1$ and $-3y + 5$?

c) $3 - 2x$ and $4 + 3y$?

E7) What is the correlation coefficient between x and y in case $n = 2$ and the two pairs of values of x and y are

a) (7.3, 4.5) and (10.4, 6.7)?

b) (4.7, 9.2) and (5.0, 6.1)?

c) (9.4, 7.0) and (10.5, 7.0)?

E8) For n observations on x, the unweighted mean is $\bar{x}$ while $\bar{x}_w$ is the weighted mean, defined by

$$\bar{x}_w = \sum_i^n w_i x_i \Big/ \sum_1^n w_i.$$

Show that $\bar{x}_w \gtreqless \bar{x}$ according as $r_{xw} \gtreqless 0$.

Hint : Find the expression for $\bar{x}_w - \bar{x}$.

E9) Find the correlation coefficient between the size of crop and the percentage of wormy fruits per apple tree for the data of E1).

In the next section, we'll talk about some algebraic relationships involving r, b_{xy} and b_{yx}.

4.6 RELATIONSHIP BETWEEN REGRESSION AND CORRELATION COEFFICIENTS

Let us first establish a connection between the regression coefficients and the correlation coefficient for given data on two variables x and y.

The regression coefficient of y on x is

$$b_{yx} = \frac{S_{xy}}{S_x^2}$$

Similarly, $b_{xy} = \frac{S_{xy}}{S_y^2}$.

Therefore, $b_{xy} \cdot b_{yx} = \frac{S_{xy}^2}{S_x^2 S_y^2} = r^2$

So, we get $|r| = \sqrt{b_{xy} \cdot b_{yx}}$ ... (18)

Note that b_{xy} and b_{yx} have the same sign. The correlation coefficient has the same sign as that of b_{xy} and b_{yx}.

Now, suppose the regression line of y on x is

$$y = 10 - 2x$$

and that of x on y is

$$x = 5 - \frac{1}{2}y.$$

Then by (18), we get $|r| = 1$.

Further, since in this case b_{xy} and b_{yx} are negative, $r = -1$. In fact, you might have noticed that, in this case, the two regression lines are identical. However, in general, the two regression lines could be quite different.

The correlation coefficient has an interesting geometric interpretation.

The regression line of x on y is given by

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$= \frac{r\sigma_x}{\sigma_y}(y - \bar{y}).$$

The regression line of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$= \frac{r\sigma_y}{\sigma_x}(x - \bar{x}),$$

where $\sigma_x^2 = \text{Var}(x)$, $\sigma_y^2 = \text{Var}(y)$ and r is the correlation coefficient. If m_1 and m_2 are the respective slopes of these two lines, then

$$m_1 = \frac{\sigma_y}{r\sigma_x} \text{ and } m_2 = \frac{r\sigma_y}{\sigma_x}$$

You can check that these lines intersect in $(\bar{x}, \bar{y})$. If we denote by θ the smaller angle between these lines, then

$$\tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2}$$

$$= \frac{\sigma_x \sigma_y (1-r^2)}{r(\sigma_x^2 + \sigma_y^2)} \quad \text{... (19)}$$

From (19), we can see that the two lines are identical (i.e., $\theta = 0$) iff $r^2 = 1$, or $r = \pm 1$. We can also infer that the only way the correlation coefficient can be zero is that the two regression lines are perpendicular to each other; one being parallel to the x-axis and the other parallel to the y-axis.

We have already seen that $r^2 = 1$ implies that $\sum_{i=1}^{n} \hat{e}_i^2 = 0$.

This, in turn, implies that each $\hat{e}_i = 0$. So, if $r = \pm 1$, then all the pairs of observations (x_i, y_i) lie on the regression line and, in such a case, y is determined by x through a mathematical relation of a straight line. Further, the value of r is + 1 if y is an increasing function of x, and it is −1, if y is a decreasing function of x.

We shall now introduce a new concept, that of product moments.

Product Moments

Recall that for univariate distributions, we have defined the rth moment about a to be

$$m'_r(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^r .$$

Extending this definition, we say that the product moment of order (r,s) of bivariate data about (a, b) is

$$m'_{rs} = \frac{1}{n}\sum_i (x_i - a)^r (y_i - b)^s, \text{ for ungrouped data,}$$

$$= \frac{1}{n}\sum_i \sum_j f_{ij}(x_i - a)^r (y_j - b)^s \text{ for grouped data.}$$

If $a = \bar{x}$ and $b = \bar{y}$, we get the central product moment of order (r, s), which we denote by m_{rs}.

Note that

$$m'_{r0} = \frac{1}{n}\sum_i (x_i - a)^r \text{ for ungrouped data}$$

$$= \frac{1}{n}\sum_i f_{i0}\,(x_i - a)^r \text{ for grouped data,}$$

where $f_{i0} = \sum_j f_{ij}$, $i = 1,2,\ldots\ldots, k$, are the marginal frequencies for x.

Similarly, we can get expressions for m_{r0}.

Thus, for $r \geq 1$, m'_{r0} and m_{r0} ignore y, and so we can say that they describe the marginal distribution of x alone. In the same way, for $s \geq 1$, m'_{0s} and m_{0s} describe the marginal distribution of y alone. In other words, the moments which characterise the joint distribution are m'_{rs} and m_{rs}, for $r \geq 1$ and $s \geq 1$.

The simplest central product moment,

$$m_{11} = \frac{1}{n}\sum (x_i - \bar{x})\,(y_i - \bar{y}) \text{ for ungrouped data}$$

$$= \frac{1}{n}\sum_i \sum_j f_{ij}\,(x_i - \bar{x})\,(y_j - \bar{y}) \text{ for grouped data.}$$

You must have found this expression familiar. We have been calling it the covariance of x and y, Cov (x, y).

Now we can write

$$r = \frac{Cov(x,y)}{s_x\, s_y} = \frac{m_{11}}{s_x s_y} \text{ and}$$

$$b_{xy} = \frac{Cov\,(x,y)}{s^2_x} = \frac{m_{11}}{s^2_x}.$$

So far, we have seen that if the scatter diagram indicates a linear relationship, then we can find the equation of the regression line and use it to predict the values of one variable when those of the other are given. The correlation coefficient tells us how precise our estimate is. But if the data do not have linear relationship, then the correlation coefficient does not serve any purpose. We shall talk about the limitations of the correlation coefficient in the next section.

4.7 LIMITATIONS OF THE CORRELATION COEFFICIENT

The correlation coefficient is a measure of the relationship between two variables, say, x and y. While it generally serves as a useful statistical tool, we should also be aware of its limitations.

Firstly, the coefficient is a measure of statistical relationship, and not of causal relationship, between the two variables. This means that the value of r tells us whether, and with what regularity, y increases or decreases as x increases. But it cannot tell us whether that increase or decrease is due to any causal (or cause-effect) relationship between the two variables. Age of husband and age of wife for a group of couples are found to be highly correlated. But to say that one of the variables is the cause of the other would be preposterous. Indeed, quite often, the high correlation between two variables is because of a third variable—a 'lurking factor', simultaneously influencing the two.

A strong correlation may often be found between two time series (i.e., series of values of two variables corresponding to, say, n points or periods of time) with no conceivable causal nexus. You will find that the correlation coefficient between the population of India and production of coal in India for the census years is very high. You will also find a high correlation coefficient between school enrolment and number of cars on the roads in the country. But this high correlation is entirely fortuitous and is due to the fact that in each case the two series are showing an increasing or decreasing trend. This type of correlation is often referred to as **nonsense (or spurious) correlation.**

Secondly, the correlation coefficient is a measure of linear statistical relationship only, and may fail to be a proper index of statistical relationship in case it is non-linear.

For example, consider the following five pairs of observations.

y	4	1	0	1	4
x	−2	−1	0	1	2

You can check that here the correlation coefficient between x and y is zero. Yet, you can see that $y_i = x_i^2$ for all i=1, 2, ..., 5. Thus, even though y and x are related by a mathematical relationship, $y = x^2$, the correlation coefficient is zero. This is because the relationship between y and x is not linear.

You should also note that a spurious correlation may be generated through the combining of non-homogeneous sets of data, i.e., sets having different means for each of the variables.

You may think of three groups of higher secondary science students: the first group is from colleges that only admit highly meritorious students, the second from colleges of the middle order where admission is less restrictive and the third from colleges where admission restrictions are virtually absent. When taken separately, each group will show a zero or near-zero correlation coefficient between score in the science subjects(x) and score in the languages(y). But in case the groups are combined, you will find a rather high (positive) correlation coefficient between x and y. This arises simply because the means of x for the three groups say, $\bar{x}_1$, $\bar{x}_2$ and $\bar{x}_3$, are unequal and also the means of y, say, $\bar{y}_1$, $\bar{y}_2$ and $\bar{y}_3$. The situation is indicated in a somewhat extreme form in Fig. 6.

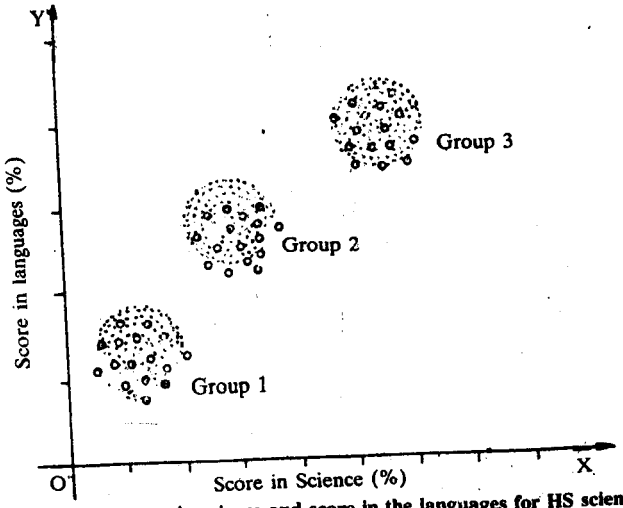


Fig. 6 : Scatter diagram for score in science and score in the languages for HS science students belonging to three distinct groups.

In the following exercise, we ask you to prove the same result algebraically.

E10) Let $\bar{x}_i$ and $\bar{y}_i$ (i=1, 2, ..., k) be the means of x and y for k groups of individuals, the ith group having n_i individuals. Suppose r_{xy}, and hence Cov(x,y), is 0 for each group. Show that for the composite group, Cov(x, y) and hence r_{xy}, will be non-zero, unless $\bar{x}_1 = \bar{x}_2 = \ldots\ldots = \bar{x}_k$ and $\bar{y}_1 = \bar{y}_2 = \ldots\ldots = \bar{y}k$.

We have seen that the correlation coefficient can be effective only when there is a linear relationship between the variables. In case of a non-linear relationship, we have to think of some other measure. In such cases, we try to fit a polynomial of degree $p > 1$ to the data. The measure of the "goodness of fit" is then provided by the correlation ratio. But we are not going to discuss non-linear regression here, as it is outside the purview of this course.

Now let us summarise what we have done in this unit.

4.8 SUMMARY

In this unit we have seen

1) how to organise bivariate data.

   We have seen how to tabulate and diagrammatically represent such data. In particular, we have seen that scatter diagrams are useful to judge what kind of relationship (if any) exists between the two variables.

2) that, in regression analysis, we try to find a mathematical model to describe the relationship between two variables.

3) how the method of least squares is used to find the equation of the regression line

   $y - \bar{y} = b_{yx}(x - \bar{x})$,

   where b_{yx} is the regression coefficient of y on x, defined by

$$b_{yx} = \frac{S_{xy}}{S_x^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

4) how to fit a regression line to given data and use it for prediction.

5) that the correlation coefficient

$$r = \frac{S_{xy}}{S_x S_y}$$

   measures the strength of the linear relationship between x and y. We have also seen how to geometrically interpret the correlation coefficient.

6) how the regression and correlation coefficients are related :

   $|r| = \sqrt{b_{xy} \cdot b_{yx}}$

   We have also defined product moments :

$$m'_{rs} = \frac{1}{n} \sum_i \sum_j f_{ij}(x_i - a)^r (y_j - b)^s ,$$

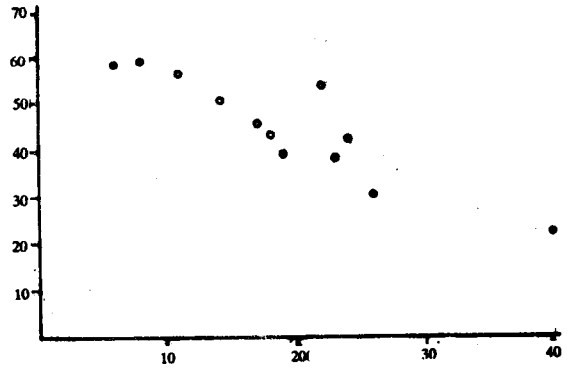
$$m_{rs} = \frac{1}{n} \sum_i \sum_j f_{ij}(x_i - \bar{x})^r (y_j - \bar{y})^s .$$

7) that the correlation coefficient suffers from certain limitations :

   - Sometimes spurious correlation is found to exist between unrelated variables.
   - The correlation coefficient measures only linear relationships.

4.9 SOLUTIONS AND ANSWERS

E1)

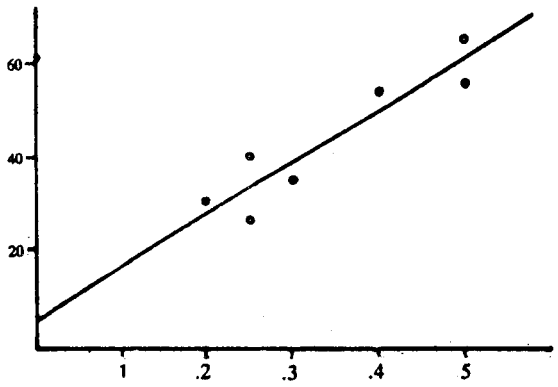


E2) $$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y}$$

$$= \sum x_i y_i - n\bar{x}\bar{y}.$$

E3) a) The dosage : independent
Reduction in blood sugar : dependent.

b)



c)

x_i	y_i	x_i^2	$x_i y_i$
.2	30	.04	6
.25	26	.0625	6.5
.25	40	.0625	10
.3	35	.09	10.5
.4	54	.16	21.6
.5	56	.25	28.0
.5	65	.25	32.5
2.4	306	0.9150	115.1

$$b_{yx} = \frac{7(115.1) - (2.4)(306)}{7(0.915) - (2.4)^2}$$

$$= \frac{71.3}{0.645}$$

$$\simeq 110.54$$

$$\bar{x} = \frac{2.4}{7} = 0.34$$

$$\bar{y} = \frac{306}{7} = 43.71$$

$\therefore a = \bar{y} - b\bar{x} \simeq 6.13$

$\therefore$ The regression line is :

$y = 6.13 + 110.54\,x.$

E4) Let A = 150, B = 600,

$$u = \frac{x - 150}{1}, \quad v = \frac{y - 600}{1}$$

u_i	v_i	u_iv_i	u_i^2
20	98	1960	400
− 3	− 82	246	9
16	125	2000	256
−25	−115	2875	625
32	145	4640	1024
−17	− 62	1054	289
− 4	−115	460	16
−25	25	− 625	625
−14	−129	1806	196
29	198	5742	841
9	88	20158	4281

$$b_{vu} = \frac{200788}{42729}$$

$\simeq 4.7$

$\therefore b_{yx} \simeq 4.7$

$\bar{y} = 600 + \bar{v} = 608.8$

$\bar{x} = 150 + \bar{u} = 150.9$

$a = -100.43.$

$\therefore$ line : $y = -100.43 + 4.7x.$

E5) $b_{yx} = \frac{d}{b} b_{uv}$, $b_{xy} = \frac{b}{d} b_{uv}$.

$\therefore r_{uv}^2 = b_{vu}\, b_{uv} = b_{yx} b_{xy} = r_{xy}^2.$

If b and d have the same sign, then b_{uv} and b_{vu} have the same sign as b_{xy} and b_{yx}. Therefore, r_{uv} has the same sign as r_{xy}. If b and d have opposite signs, by the same argument, r_{uv} and r_{xy} have opposite signs.

E6) a) + 0.73, b) − 0.73 c) − 0.73

E7) Since only a pair of observations are given, they clearly fall on a line.

a) r = + 1 (y increases as x increases)

b) r = − 1

c) r is not defined, since $s_y = 0$.

E8) $$\bar{x}_w - \bar{x} = \frac{1}{\sum w_i}\left[\sum_i w_i x_i - \left(\sum_i w_i\right)\left(\sum_i x_i\right)/n\right]$$

$$= \frac{n\,\mathrm{Cov}\,(w, x)}{\sum w_i} \gtreqless 0,$$

according as Cov(w,x) $\gtreqless$ 0, since $w_i \geq 0\ \forall\ i$.

E9)

x_i	y_i	x_i^2	y_i^2	x_iy_i
8	59	64	3481	472
6	58	36	3364	348
11	56	121	3136	616
22	53	484	2809	1166
14	50	196	2500	700
17	45	289	2025	765
18	43	324	1849	774
24	42	576	1764	1008
19	39	361	1521	741
23	38	529	1444	874
26	30	676	900	780
40	20	1600	400	800
228	533	5256	25193	9044

$$r = \frac{9044 \times 12 - 228 \times 533}{\sqrt{12 \times 5256 - (228)^2}\ \sqrt{12 \times 25193 - (533)^2}}$$

$$= \frac{-12996}{105.3 \times 135}$$

$$= -0.914.$$

E10) Let s_{xy} be the covariance in the composite group and s_{xy}^i, the covariance in the ith group. We have

$$s_{xy} = \frac{1}{n}\left\{\sum_1^k n_i s_{xy}^i + \sum_1^k n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})\right\},$$

where $n = \sum_1^k n_i$ and $\bar{x}$, the composite mean of x, $\bar{y}$, the composite mean of y.

$$\therefore s_{xy} = \frac{1}{n}\sum n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) .$$

Therefore, the result follows.

NOTES

NOTES

Block

2

PROBABILITY ON DISCRETE SAMPLE SPACES

Course Design Committee

Prof. S.K. Mitra (*Chairman*)
Indian Statistical Institute
New Delhi

Prof. A.M. Goon
Presidency College
Calcutta

Prof. J. Medhi
Guwahati

Prof. B.L.S. Prakasa Rao
Indian Statistical Institute
New Delhi

Prof. Aloke Dey
Indian Statistical Institute
New Delhi

Prof. K. Balasubramanian
Indian Statistical Institute
New Delhi

Prof. D.D. Joshi
Ex-Pro-Vice-Chancellor
IGNOU

Dr. V. Madan
School of Sciences
IGNOU

Dr. Poornima Mital
School of Sciences
IGNOU

Dr. Manik Patwardhan
School of Sciences
IGNOU

Dr. Sujatha Varma
School of Sciences
IGNOU

Block Preparation Team

Prof. S.K. Mitra (*Editor*)
ISI, New Delhi

Prof. Aloke Dey (*Co-editor*)
ISI, New Delhi

Prof. S.R. Adke
University of Poona
Pune

Dr. Manik Patwardhan
School of Sciences
IGNOU

Course Coordinator : Dr. Manik Patwardhan

Acknowledgement
To Prof. R.K. Bose, Dr. Parvin Sinclair and Dr. Sujatha Varma for their useful comments on the manuscript.

Production

Mr. Balakrishna Selvaraj
Registrar (PPD)
IGNOU

Mr. M.P. Sharma
Joint Registrar (PPD)
IGNOU

March, 1993

ISBN-81-7263-304-1

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by Mr. Balakrishna Selvaraj, Registrar (PPD).

Notations and Symbols

Ω	:	Sample Space
$P\{\omega\}$	:	Probability of ω
$P(A)$	:	Probability of the event A
$P(A \mid H)$	:	Conditional probability of A, given that H has occurred
X, Y, X_i	:	Random variables
$[X = j]$	:	$\{\omega \in \Omega \mid x(\omega) = j\}$
$P[X = j]$	:	Probability that X equals j
p.m.f.	:	Probability mass function
$f(x_j)$	:	$P[X = x_j]$
b(j, n, p)	:	P[X = j], where X is a binomial variable with parameters, (n, p).
h(j; n, N, M)	:	P[X = j], where X has a hypergeometric distribution with parameters, (n, N, M).
$\binom{\alpha}{j}$	:	$\frac{\alpha(\alpha-1)\ldots(\alpha-j+1)}{j!}$, where $-\infty < \alpha < \infty$, j is a non-negative integer.
P(r, m)	:	P[X = r], where X has a Poisson distribution with parameter $m > 0$.

Also see the list in Block 1.

BLOCK 2 PROBABILITY ON DISCRETE SAMPLE SPACES

You have studied methods of representing and summarising statistical data of different types in Block 1. The purpose of the techniques that you studied there, is to help in the understanding and simplification of the information contained in data sets. The next stage in this exercise is to try to understand the pattern and causes of observed variability. We can then construct techniques of reaching objective conclusions about the population on the basis of information provided by the sample. This is possible only with the help of probability theory which we introduce in this block for the simplest of situations.

A French nobleman Chevalier de Mere, who had considerable experience in gambling, noticed some contradictions between his theoretical conclusions and the observed results of games of chance. He discussed his doubts with Pascal (1623-1662) who was a famous mathematician of the time. Pascal solved de Mere's difficulties and a few more problems. On hearing about these problems from Pascal, another mathematician, Fermat (1601-1665), became interested in them. Pascal and Fermat corresponded with each other and laid the foundations of the theory of probability. It was mainly in Europe during the seventeenth and eighteenth century that mathematicians like Huygens (1629-1695), Jacob Bernoulli. (1654-1705), de Moivre (1667-1754), Laplace (1749-1827) and Poisson (1781-1840) made important contributions and developed the theory of probability into a distinct new branch of mathematics.

After this early development of probability theory in Europe, there developed in Russia a very strong and important school of studies in Probability. The Russian mathematicians who made substantial and lasting contributions to probability theory are Chebychev (1821-1894), his students A. Markov (1856-1922) and Liapounov (1868-1918), S. Bernstein, and A. Khintchine.

The subject has been put on a sound mathematical base only in the third and fourth decades of this century, mainly due to the pioneering work of the great Russian mathematician A.N. Kolmogorov (1903-1987).

In Unit 5 we shall analyse the nature of experiments whose results cannot be predicted in advance. We shall explore the definition of the probability of an event and its simple properties in Unit 6. Unit 7 deals with the concept of a random variable and its probability distribution. The most commonly used probability distributions are discussed in Units 8 and 9.

The mathematical background needed for this block is a knowledge of permutations and combinations, an ability to undertake algebraic simplifications and logical thinking. Although pen, paper and patience are sufficient for this block, a good hand-held scientific calculator would be an asset to solve numerical problems.

UNIT 5 SAMPLE SPACE OF A RANDOM EXPERIMENT

Structure

5.1 INTRODUCTION

Many situations arise in our everyday life as well as in scientific, administrative or organisational work, where we cannot predict the outcome of our actions or of the experiment we are conducting. Such experiments, whose outcome cannot be predicted, are called random experiments. We give a wide variety of examples in Sec. 5.2 to explain the concept of a random experiment. The set of all possible outcomes of an experiment is called its sample space. We have illustrated the different types of sample spaces that we generally come across in Sec. 5.3. Section 5.4 deals with the study of events associated with a random experiment whose sample space is either finite or countably infinite. In Sec. 5.5 we discuss methods of combining events to generate new events. Here is a list of what you should be able to do by the end of this unit.

Objectives

After studying this unit you should be able to :

- distinguish between random and non-random experiments,
- specify the sample space of a random experiment and classify it as discrete or continuous,
- identify events with subsets of the sample space,
- examine and identify relations between events,
- generate new events out of a given collection of events.

5.2 RANDOM EXPERIMENTS

We give below some examples of a random experiment :

- A physicist performs an experiment to discover laws governing the flow of an electrical current or the propagation of sound, heat or light etc.
- A chemist studies the reactions of chemicals and tries to understand the chemical properties of matter.
- A physician compares two or more drugs to find out the most effective one by trying them out on experimental animals or on patients.
- To describe the relationship between the price of a commodity and its demand and supply, an economist observes the values assumed by these variables by conducting a market survey over a period of time.

With a little imagination, we can construct many more examples of such experiments.

Experimentation is not necessarily restricted to a laboratory or to a university or a college. It forms an important part of our everyday life. When you buy a dress or a shirt, when you

vote for a candidate at an election, when you inspect a few grains of rice to decide whether the rice is cooked or not, when you decide to register for this course, you are performing an experiment. Thus, experimentation constitutes an integral part of our lives as well as our learning processes. In this unit we shall develop methods of describing the results of an experiment. Once we can describe the results we'll be able to talk about the chances of their occurrence.

Consider the following simple experiments :

Experiment 1 : A stone is allowed to fall freely from height and we observe whether or not the stone hits the ground.

Experiment 2 : Water in a pot is heated for a sufficiently long time to a temperature greater than 100°C. We observe whether the water turns into steam.

In these experiments, we have no doubt about the final outcome. The stone will eventually hit the ground. The water in the pot will ultimately turn into steam. These experiments have only one possible outcome. Even if these experiments are repeated again and again, every such repetition will yield the same result.

On the other hand, in the following experiments there are two or more possible results.

Experiment 3 : A coin is tossed to decide which of the two teams A and B would bat first in a game of cricket. The coin may turn up a head or a tail.

Experiment 4 : A person coming out of a polling centre is requested to disclose the name of the candidate in whose favour he/she has voted. He/she may refuse to tell us or give the name of any one of the candidate.

Experiment 5 : Three consecutive items produced by a machine are inspected and classified as good or bad (defective). We may get 0, 1, 2, or 3 defective items as a result of this inspection.

Experiment 6 : A newly invented vaccine against a disease is given to 30 healthy people. These thirty people as well as another group of 20 similar people who are not vaccinated, are watched over the next six months to see whether they develop the disease. The total number of affected people may vary between 0 and 50.

Experiment 7 : A small town has 100 telephones. The number of busy telephones between 9 and 10 a.m. is noted for each day of a week. The number of busy telephones may be any number between 0 to 100.

Experiment 8 : A group of ten persons is classified according to their blood groups O, A, B and AB. The number of persons in each group may vary between 0 and 10, subject to the frequencies of all four classes adding up to 10.

Experiment 9 : The number of accidents along the Bombay-Bangalore national highway during the month is noted.

Experiment 10 : A radio-active substance emits particles called α-particles. The number of such particles reaching an observation screen during one hour is noted.

Experiment 11 : Thirteen cards are selected without replacement from a well-shuffled pack of 52 playing cards.

The nine experiments, 3-11, have two common features.

i) Each of these experiments have more than one possible outcome.

ii) It is impossible to predict the outcome of the experiment.

For example, we cannot predict whether a coin, when it is tossed, will turn up a head or a tail (Experiment 3). Can we predict without error the number of busy telephones (Experiment 7)? It is impossible to predict the 13 cards we shall obtain from a well-shuffled pack (Experiment 11).

Do you agree that all the experiments 3-11 have the above-mentioned features (i) and (ii)? Go through them carefully again, and convince yourself.

This discussion leads us to the following definition.

Definition 1 : An experiment with more than one possible outcome and whose result cannot be predicted, is called a **random experiment.**

Sample Space of a Random Experiment

So, Experiments 3 to 11 are random experiments, while in Experiments 1 and 2 the outcome of the experiment can be predicted. Therefore, Experiments 1 and 2 do not qualify as random experiments. You will meet many more illustrations of random experiments in this and subsequent units.

You may now try this exercise.

E1) Classify the experiments described below as random or non-random experiments.

a) A spark of electricity is introduced in a cylinder containing a mixture of hydrogen and oxygen. The end product is observed.

b) A lake contains two types of fish. Ten fish are caught and the number of fish of each type is noted.

c) The time taken by a powerful radio impulse to travel from the earth to the moon and for its echo to return to the sender is observed.

d) Two cards are drawn from a well-shuffled pack of 52 playing cards and the suits (Club, Diamond, Heart and Spade) to which they belong are noted.

In the dictionary you will find that something that is **random**, happens or is chosen without a definite plan, pattern or purpose.

In the next section we shall talk about the set of all possible outcomes of a random experiment.

5.3 SAMPLE SPACE

In the previous section you have seen a number of examples of random experiments. The first step we take in the study of such experiments is to specify the set of all possible outcomes of the experiment under consideration.

When a coin is tossed (Experiment 3), either a head turns up or a tail turns up. We do not consider the possibility of the coin standing on its edge or that of its rolling away out of sight. Thus, the set Ω of all possible outcomes consists of two elements, Head and Tail. Therefore, we write Ω = {Head, Tail} or, more simply, Ω = {H, T}.

Ω is the Greek letter capital 'omega'

In Experiment 4, the person coming out of the polling centre may give us the name of the candidate for whom he/she voted, or may refuse to disclose his/her choice. If there are 5 candidates C_1, C_2, C_3, C_4 and C_5, seeking election, then there are six possible outcomes, five corresponding to the five candidates and the sixth one corresponding to the refusal R of the interviewed person to disclose his/her choice. The set of all possible outcomes is thus, $\Omega = \{C_1, C_2, C_3, C_4, C_5, R\}$.

Note that here we have ignored certain possibilities, like the possibility of the person not voting at all or voting in such a manner that his/her ballot paper becomes invalid.

Experiment 5 is comparatively simple, if we agree that it is possible to classify each item as Good (G) or Bad (B) without error. Then Ω = {GGG, GGB, GBG, BGG, BBG, BGB, GBB, BBB} where, for example, GBG denotes the outcome when the first and third units are good and the second one is bad.

The situation in Experiment 6 is a little more complicated. To test the efficacy of the vaccine, we will have to look at the number of vaccinated persons who were affected (x) and the number of non-vaccinated ones who were affected (y). Here x can be any integer between 0 and 30 and y can be any integer between 0 and 20. The set Ω of all possible outcomes is

$$\Omega = \{(x, y) \mid x = 0, 1, \ldots, 30, y = 0, 1, 2, \ldots, 20\}.$$

This specification of Ω is valid only if we assume that we are able to observe all the 50 persons for the entire period of six months. In particular, we assume that none of them becomes untraceable because of his/her leaving the town or because of his/her death due to some other cause.

In the illustrations discussed so far, do you notice that the number of points in Ω is finite in each case? It is 2 for Experiment 3, 6 for Experiment 4, $31 \times 21 = 651$ for Experiment 6. But this is not always true.

Consider, for example, Experiments 9 and 10. The number of accidents along the Bombay-Bangalore highway during the month of observation can be zero, one, two, . . . or some other positive integer. Similarly, the number of α-particles emitted by the radio-active substance can be any positive integer. Can we say that the number of accidents or α-particles would not exceed a specified limit? No. Because of this, and also in order to simplify our mathematics, we usually postulate that in both these examples the set of all possible outcomes is $\Omega = \{0, 1, 2, \ldots\}$, i.e., it is the set of all non-negative integers. We are now in a position to introduce certain terms in a formal manner.

Definition 2 : The set Ω of all possible outcomes of an experiment E is called the **sample space** of the experiment. Each individual outcome of E is called a **point,** a **sample point** or an **element** of Ω.

You would also notice that in every experiment that was discussed, we made certain assumptions like the coin not being able to stand on its edge or not rolling away, all the fifty persons being available for the entire period of six months for observation, etc. Such assumptions are necessary to simplify our problems as well as our mathematics.

In all the examples discussed so far, the sample space is either a finite set, i.e., a set containing a finite number of points or is an infinite set whose elements can be arranged in an unending sequence, i.e., has a countable infinity of elements. We have a special name for such spaces.

Definition 3 : A sample space containing a finite number of points or a countable infinity of points is called a **discrete sample space.**

In this block we shall be concerned only with discrete sample spaces. However, there are many situations where we have to deal with sample spaces which are not discrete. For example, consider the age of a person. Although there are limitations to the accuracy with which we can measure the age of a person, in the idealised situation we can think of age being any number between 0 and ∞. Of course, no one has met a person with infinite age of for that matter who is more than 150 years old. Nevertheless, most of the actuarial and demographic studies are carried out assuming that there is no upper bound on age. Thus, we may say that the sample space of the experiment of finding out the age of an arbitrarily selected person is the interval $]0, \infty[$. Since the elements of the interval $]0, \infty[$ cannot be arranged in a sequence, such a sample space is not a discrete sample space.

Some other examples where non-discrete sample spaces are appropriate are (i) the price of wheat, (ii) the amount of ozone in a volume of space, (iii) the length of a telephone conversation, (iv) the duration one spends in a queue, (v) the yield of rice in our country in one year.

In all these examples, it is necessary to deal with non-discrete sample spaces, However, we'll defer the study of probability theory for such experiments to the next block.

Now see if you can solve this exercise.

E2) Write down the sample spaces of all those experiments from 3 to 11 which we have not discussed earlier. Indicate in each case the assumptions made by you.

Now that we have seen how to specify the elements of a sample space, we can talk about the events associated with it.

5.4 EVENTS

We have described a number of random experiments till now. We have also identified the sample spaces associated with them. In the study of random experiments, we are interested not only in the individual outcomes but also in certain events. As you will see later, events are subsets of the sample space. In this section we shall formalise the intuitive concept of an event associated with a random experiment which has a discrete sample space. We shall also

study methods of generating new events from specified ones and study their inter-relationships.

Sample Space of a Random Experiment

Consider the experiment of inspecting three items (Experiment 5). The sample space has the eight points,

GGG, GGB, GBG, BGG, BBG, BGB, GBB, BBB.

We label these points $\omega_1, \omega_2, \ldots, \omega_8$, respectively.

Suppose we are interested in those outcomes which correspond to the event of obtaining exactly one good item in the three inspected items. The corresponding sample points are ω_5 = BBG, ω_6 = BGB and ω_7 = GBB. Thus, the subset $\{\omega_5, \omega_6, \omega_7\}$ of the sample space corresponds to the "event" A that only one of the inspected items is good.

ω is the lower case Greek letter 'omega'.

On the other hand, consider the subset C = $\{\omega_5, \omega_6, \omega_7, \omega_8\}$ consisting of the points BBG, BGB, GBB, BBB. We can identify the subset C with the event "There are at least two bad items."

This discussion suggests that we can associate a subset of the sample space with an event and an event with a subset. This leads us to the following definition.

Definition 4 : When the sample space of an experiment is discrete, any subset of the sample space is called an **event.**

Thus, we also consider the empty set as an event.

You will soon find that the two extreme events, φ and Ω, consisting, respectively, of no points and all the points of Ω are most uninteresting. But we need them to complete our description of the class of all events. In fact, φ is called the **impossible event** and Ω is called the **sure event**, for reasons which will be obvious in the next unit. Also, note that an individual outcome ω, when identified with the singleton (ω), constitutes an event.

The following example will help you in understanding events.

Example 1 : Suppose we toss a coin twice. The sample space of this experiment is Ω = {HH, HT, TH, TT}, where HT stands for a head followed by a tail, and other points are similarly defined. Let's list all the events associated with this experiment. There are 16 such events. These are :

φ, {HH}, {HT}, {TH}, {TT}

{HH, HT}, {HH, TH}, {HH, TT}, {HT, TH}

{HH, TT}, {TH, TT}, {HH, HT, TH}, {HH, TH, TT},

{HH, HT, TT}, {HT, TH, TT}, Ω.

Since we have identified an event with a subset of Ω, the class of all events is the class of all the subsets of Ω. If Ω has N points, for a fixed r, we can form $\binom{N}{r}$ sets consisting of r points, where r = 0, 1, . . . , N. The total number of events is, therefore,

$$\binom{N}{0} + \binom{N}{1} + \ldots + \binom{N}{N} = (1+1)^N = 2^N.$$

By binomial theorem
$(1+x)^N = \binom{N}{0} + \binom{N}{1}x + \ldots + \binom{N}{N}x^N.$

In Example 1, N = 4. Therefore, we have $2^4 = 16$ events. If N = 10, we shall $2^{10} = 1024$ events. The number of events thus increases rapidly with N. It is infinite if the sample space is infinite.

Let us now clarify the meaning of the phrase "The event A has occurred."

We continue with Experiment 5. Let A denote the event $\{\omega_5, \omega_5, \omega_7\}$ = {BBG, BGB, GBB}. If, after performing the experiment, our outcome is ω_5 = BBG, which is a point of the set A, we say that the event A has occurred. If, on the other hand, the outcome is ω_8 = BBB, which is not a point of A, then we say that A has not occurred. In other words, given the outcome ω of the experiment, we say that A has occurred if $\omega \in$ A and that A has not occurred if $\omega \notin$ A.

On the other hand, if we only know that A has occurred, all we know is that the outcome of

the experiment is one of the points of A. It is then not possible to decide which individual outcome has resulted unless A is a singleton.

In the next section we shall talk about some ways of combining events.

5.5 ALGEBRA OF EVENTS

In this section we shall study different ways in which we can combine two or more events. We shall also study relations between them. Since we are dealing with discrete sample spaces and since any subset of the sample space is an event, we shall use the terms event and subset interchangeable.

In what follows, events and sets are denoted by capital letters A, B, C, . . . , with or without suffixes. We shall assume that they all consist of points chosen from the same sample space Ω.

Let Ω = {GGG, GGB, GBG, BGG, BBG, BGB, GBB, BBB} be the sample space corresponding to Experiment 5. Let A = {BBG, BGB, GBB} be the event that only one of the three inspected items is good. Here the point BGB is an element of the set A and the point BBB is not an element of A. We express this by writing BGB $\in$ A and BBB $\notin$ A.

Recall, (Unit 1, MTE-04) that the complement, $A^c = \{\omega \in \Omega | \omega \notin A\}$. Then $\phi^c = \Omega$ and $\Omega^c = \phi$. Fig. 1 shows a Venn diagram representing the sets A and A^c.

Suppose, now, that the outcome of the experiment is BBB. Obviously, the event A has not occurred. But, we may say the event "not A" has occurred. In probability theory, the event "not A" is called the **event complementary to A** and is denoted by A^c.

Let's try to understand this concept by looking back at Experiments 3-11.

Example 2

i) For Experiment 5, if A = {BBG, BGB, GBB}, then

$$A^c = \{GGG, GGB, BGG, GBG, BBB\}.$$

ii) In Experiment 6, let A denote the event that the number of infected persons is at most 40. Then

$$A^c = \{(x, y) \mid x + y > 40, x = 0, 1, \ldots, 30, y = 0, 1, \ldots, 20\}.$$

iii) In Experiment 11, if B denotes the event that none of the 13 cards is a spade, B^c consists of all hands of 13 cards, each one of which has at least one spade.

Suppose now that A_1 and A_2 are two events associated with an experiment. We can get two new events, $A_1 \cap A_2$ (A_1 intersection A_2) and $A_1 \cup A_2$ (A_1 union A_2) from these two. With your knowledge of set theory (MTE-04), you would expect the event $A_1 \cap A_2$ to correspond to the set whose elements belong to both A_1 and A_2. Thus,

$$A_1 \cap A_2 = \{\omega | \omega \in A_1 \text{ and } \omega \in A_2\}.$$

Similarly, the event $A_1 \cup A_2$ corresponds to the set whose elements belong to at least one of A_1 and A_2.

$$A_1 \cup A_2 = \{\omega | \omega \in A_1 \text{ or } \omega \in A_2\}.$$

Fig. 2 (a) and (b) show the Venn diagrams representing $A_1 \cap A_2$ and $A_1 \cup A_2$.

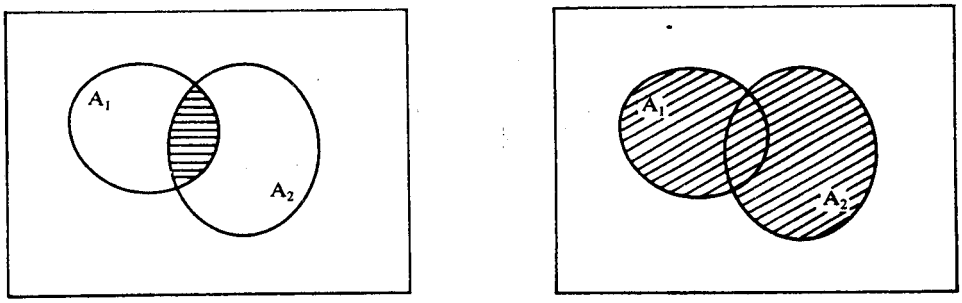


Fig. 2 : The shaded region represents the set (a) $A_1 \cap A_2$ (b) $A_1 \cup A_2$.

We'll try to clarify this concept with some examples.

Example 3 : In many games of chance, a small cube (or die) with equal sides, bearing numbers 1, 2, 3, 4, 5, 6, or dots 1-6 on its six faces (Fig. 3), is used. When such a symmetric die is thrown, one of its six faces would be uppermost. The number (or number of dots) on the uppermost faces is called the score obtained on the throw or roll of a die. The appropriate sample space for the experiment of throwing a die is then $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let A_1 be the event that the score exceeds three and A_2 be the event that the score is even. Then

$$A_1 = \{4, 5, 6\}, A_2 = \{2, 4, 6\}$$

Therefore, $A_1 \cap A_2 = \{4, 6\}$ and

$$A_1 \cup A_2 = \{2, 4, 5, 6\}.$$

Suppose now that the score is 6. We can say that A_1 has occurred. But then A_2 has also occurred. In other words, both A_1 and A_2 have occurred. Thus, the simultaneous occurrence of A_1 and A_2 corresponds to the occurrence of the event $A_1 \cap A_2$.

When the outcome is 5, A_1 has occurred but A_2 has not occurred. Further, when the outcome is 2, A_2 has occurred and A_1 has not. When the outcome is 4, both A_1 and A_2 have occurred. In case of each of these outcomes, 2, 5 or 4, we notice that at least one of A_1 and A_2 has occurred. Note, further, that $A_1 \cup A_2$ has also occurred. Thus, the occurrence of at least one of the two events A_1 and A_2 corresponds to the occurrence of $A_1 \cup A_2$.

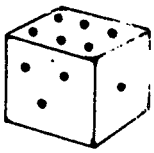


Fig. 3

Example 4 : Suppose the die in Example 3 is thrown twice. Then Ω is the set $\{(x, y) \mid x, y = 1, 2, 3, \ldots, 6\}$ consisting of thirty-six points (x, y), where x is the score obtained on the first throw and y, that obtained on the second throw. If B_1 is the event that the score on the first throw is six and B_2 the event that the sum of the two scores is at least 11, then

$$B_1 = \{(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$$

and

$$B_2 = \{(5, 6), (6, 5), (6, 6)\}.$$

What are $B_1 \cap B_2$ and $B_1 \cup B_2$? You can check that

$$B_1 \cap B_2 = \{(6, 5), (6, 6)\}$$

and

$$B_1 \cup B_2 = \{(5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}.$$

The union and intersection of two sets can be utilised to define union and intersection of three or more sets.

So, if $A_1, A_2, \ldots, A_n$ are n events, then we define

$$\bigcap_{j=1}^{n} A_j = \{\omega \mid \omega \in A_j \text{ for every } j = 1, \ldots, n\}$$

and

$$\bigcup_{j=1}^{n} A_j = \{\omega \mid \omega \in A_j \text{ for at least one } j = 1, \ldots, n\}.$$

Note that the occurrence of $\bigcap_{j=1}^{n} A_j$ corresponds to the simultaneous occurrence of all the n events and the occurrence of $\bigcup_{j=1}^{n} A_j$ corresponds to that of at least one of the n events $A_1, \ldots, A_n$. We can similarly define the union and intersection of an infinite number of events, $A_1, A_2, \ldots, A_n, \ldots$.

Another set operation with which you are familiar is a combination of complementation and intersection. Let A and B be two sets. Then the set $A \cap B^c$ is usually called the difference of A and B and is denoted by A – B. It consists of all points which belong to A but not to B.

Thus, in Example 4,

$$B_1 - B_2 = \{(6, 1), (6, 2), (6, 3), (6, 4)\}$$

and

$$B_2 - B_1 = \{(5, 6)\}$$

In this notation, A^c is the set $\Omega - A$. You can see the Venn diagram for A – B in Fig. 4.

Now, suppose A_1, A_2 and A_3 are three arbitrary events. What does the occurrence of $A_1 \cap A_2^c \cap A_3^c$ signify?

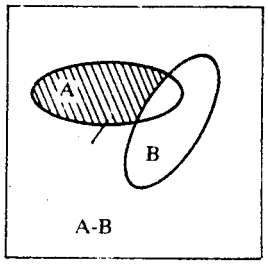


Fig. 4 : The shaded portion represents A - B.

This event occurs iff only A_1 out of A_1, A_2 and A_3 occurs, that is, iff A_1 occurs but neither A_2 nor A_3 occur.

If you have followed this, you should be able to do this exercise quite easily.

E3) If A_1, A_2 and A_3 are three arbitrary events, what does the occurrence of the following events signify?

a) $E_1 = A_1 \cap A_2 \cap A_3$

b) $E_2 = A_1^c \cap A_2^c \cap A_3^c$

c) $E_3 = (A_1 \cap A_2 \cap A_3^c) \cup (A_1 \cap A_3 \cap A_2^c) \cup (A_2 \cap A_3 \cap A_1^c)$

d) $E_1 \cup E_3$

The set operations like formation of intersection, union and complementation of two or more sets that we have listed above and their combinations are sufficient for constructing new events out of old ones. However, we need to express in a precise way commonly used expressions like (i) if the event A has occurred, B could not have occurred and (ii) the occurrence of A implies that of B. We'll explain this by taking an example first.

Example 5 : Let us consider the following experiments.

i) In the experiment of tossing a die twice, let A be the event that the total score is 8 and B that the absolute difference of the two scores is 3. Then

$$A = \{(x, y) \mid x + y = 8, x, y = 1, 2, 3, \ldots, 6\}$$
$$= \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$$

and $B = \{(x, y) \mid |x - y| = 3, x, y = 1, 2, 3, \ldots, 6\}$
$$= \{(1, 4), (2, 5), (3, 6), (6, 3), (5, 2), (4, 1)\}.$$

ii) Consider Experiment 11, where we select 13 cards without replacement from a pack of cards. Let

event A : all the 13 cards are black and

event B : there are 6 diamonds and 7 hearts.

Note that in both the cases there is no point which is common to both A and B. Or in other words, $A \cap B$ is the empty set. Therefore, in both i) and ii) we conclude that if A occurs, B cannot occur and conversely, if B occurs A cannot occur.

Now let us find an example for the situation : the occurrence of A implies that of B. Take the experiment of tossing a die twice. Let $A = \{(x, y) \mid x + y = 12\}$ be the event that the total score is 12, and $B = \{(x, y) \mid x - y = 0\}$ be the event of having the same score on both the throws. Then

$$A = \{(6, 6)\} \text{ and}$$

$$B = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\};$$

so that whenever A occurs, B does. Note that $A \subset B$.

You were already familiar with the various operations on sets. In Sec. 5.4 we had identified events with subsets of the sample space. What we have done in this section is to apply set operations to events, and to interpret the combined events.

You can check whether you have grasped these ideas by doing the following exercises.

E4) Let A_1, A_2, A_3 and A_4 be arbitrary events. Find expressions for the events that correspond to occurrence of

a) only A_1 and A_2,

b) none of A_1, A_2, A_3 and A_4,

c) one and only one of A_1, A_2, A_3, A_4,

d) not more than one of A_1, A_2, A_3, A_4,

e) at least two of A_1, A_2, A_3, A_4.

E5) Express in words the following events :

a) $A_1^c \cap A_2 \cap A_3$

b) $(A_1^c \cap A_2^c \cap A_3^c) \cup (A_1 \cap A_2^c \cap A_3^c)$

c) $(A_1 \cup A_2) - (A_3 \cup A_4)$

d) $(A_1 \cup A_2) \cap A_3$

e) $(A_1 \cap A_2) \cap (A_2 \cap A_3) \cup (A_3 \cap A_1)$

Now, before ending this unit let us go over its main points.

5.6 SUMMARY

In this introductory unit to the study of probability, we have made the following points :

1) There are many situations in real life as well as in scientific work which can be regarded as experiments having more than one possible outcome. We cannot predict the outcome that we will obtain at the conclusion of the experiment. Such experiments are called **random experiments.**

2) The study of random experiments begins with a specification of its all possible outcomes. In this specification, we have to make certain assumptions to avoid complexities. The set of all possible outcome is called the **sample space** of the experiment. A sample space with a finite number or a countable infinity of points is a **discrete** sample space.

3) When we are dealing with a discrete sample space, we can identify events with sets of points in the sample space. Thus, an event can be formally regarded as a subset of the sample space. This definition works only when the sample space is discrete.

4) We can use operations like complementation, intersection, union and difference to generate new events.

5) Some complex events can be described in terms of simpler events by using the above-mentioned set operations.

5.7 SOLUTIONS AND ANSWERS

E1) a) This is a non-random experiment as the electrical spark would ignite the hydrogen and it would combine with oxygen to produce water.

b) This is a random experiment as one cannot predict the number of fish of each type that would be caught.

c) The radio impulse travels with the velocity of light which is known to be

a physical constant. The time for the radio impulse to reach the moon and for its echo to return can be predicted without error. Hence this is a non-random experiment.

d) A random experiment.

E2) The sample space for the random experiment described in

i) Experiment 7 is $\Omega = \{0, 1, \ldots, 100\}$

ii) Experiment 8, is

$\Omega = \{(x_0, x_A, x_B, x_{AB}) \mid x_0 + x_A + x_B + x_{AB} = 10\}$.

where x_0, x_A, x_B and x_{AB} are the number of persons with blood-groups 0, A, B and AB, respectively, in the group of 10 persons.

iii) Experiment 11, is the set of all possible, i.e.,

$\binom{52}{13} = 2.476552 \times 10^{13}$, suits of 13 cards that can be formed out of 52 cards.

E3) a) The event $E_1 = A_1 \cap A_2 \cap A_3$ occurs if all the three occur.

b) None of the three events A_1, A_2, A_3 occurs iff $A_1^c \cap A_2^c \cap A_3^c$ occurs.

c) The event E_3 occurs if exactly two of the three events occur.

d) $E_1 \cup E_3$ corresponds to occurrence of at least two of the three events A_1, A_2 and A_3.

E4) a) $A_1 \cap A_2 \cap A_3^c \cap A_4^c$

b) $(A_1^c \cap A_2^c \cap A_3^c \cap A_4^c) = E_1$, say.

c) $(A_1 \cap A_2^c \cap A_3^c \cap A_4^c) \cup (A_1^c \cap A_2 \cap A_3^c \cap A_4^c) \cup (A_1^c \cap A_2^c) \cap (A_3 \cap A_4^c)$ $\cup A_1^c \cap A_2^c \cap A_3^c \cap A_4) = E_2$, say.

d) $E_1 \cup E_2$

e) $(E_1 \cup E_2)^c$.

E5) These events corresponds to occurrence of

a) A_2 and A_3 but not A_1,

b) None of A_2 and A_3,

c) At least one of A_1 and A_2, but none of A_3 and A_4,

d) A_3 and at least one of A_1 and A_2,

e) At least two out of A_1, A_2 and A_3.

UNIT 6 PROBABILITY ON A DISCRETE SAMPLE SPACE

Structure

6.1 INTRODUCTION

In this unit, we shall introduce you to some simple properties of the probability of an event associated with a discrete sample space. Our definitions require you to first specify the probabilities to be attached to each individual outcome of the random experiment. Therefore, we need to answer the question : How does one assign probabilities to each and every individual outcome? This question was answered very simply by the classical probabilists (like Jacob Bernoulli). They **assumed** that all outcomes are equally likely. Therefore, for them, when a random experiment has a finite number N of outcomes, the probability of each outcome would be 1/N. Based on this assumption they developed a probability theory, which we shall briefly describe in Sec. 6.4. However, this approach has a number of logical difficulties. One of them is to find a reasonable way of specifying "equally likely outcomes."

However, one possible way out of this difficulty is to relate the probability of an event to the relative frequency with which it occurs. To illustrate this point, we consider the experiment of tossing a coin a large number of times and noting the number of times "Head" appears. In fact, the famous mathematician, Karl Pearson, performed this experiment 24000 times. He found that the relative frequency, which is the number of heads divided by the total number of tosses, approaches 1/2 as more and more repetitions of the experiment are performed. This is the same figure which the classical probabilists would assign to the probability of obtaining a head on the toss of a balanced coin.

Thus, it appears that the probability of an event could be interpreted as the long range relative frequency with which it occurs. This is called the statistical interpretation or the frequentist approach to the interpretation of the probability of an event. This approach has its own difficulties. We'll not discuss these here. Apart from these two, there are a few other approaches to the interpretation of probability. These issues are full of philosophical controversies, which are still not settled.

We, shall adopt the axiomatic approach formulated by Kolmogorov and treat probabilities as numbers satisfying certain basic rules. This approach is introduced in Sec. 6.2.

In Sec. 6.2 and 6.3 we deal with properties of probabilities of events and their computation. We discuss the important concept of conditional probability of an event given that another event has occurred in Sec. 6.4. It also includes the celebrated Bayes' theorem. In Sec. 6.5 we discuss the definition and consequences of the independence of two or more events. Finally, we talk about the probabilistic structure of independent repetitions of experiments in Sec. 6.6. After getting familiar with the computation of probabilities in this unit, we shall take up the study of probability distributions in the next one.

Objectives

After studying this unit you should be able to :

- assign probabilities to the outcomes of a random experiment with discrete sample space,
- establish properties of probabilities of events,
- calculate the probability of an event,
- calculate conditional probabilities and establish Bayes theorem,
- check and utilise the independence of two or more events.

6.2 PROBABILITY : AXIOMATIC APPROACH

We have considered a number of examples of random experiments in the last unit. The outcomes of such experiments cannot be predicted in advance. Nevertheless, we frequently make vague statements about the chances or probabilities associated with outcomes of random experiments. Consider the following examples of such vague statements :

i) It is very likely that it would rain today.

ii) The chance that the Indian team will win this match is very small.

iii) A person who smokes more than 10 cigarettes a day will most probably developing lung cancer.

iv) The chances of my winning the first prize in a lottery are negligible.

v) The price of sugar would most probably increase next week.

Probability theory attempts to quantify such vague statements about the chances being good or bad, small or large. To give you an idea of such quantification, we describe two simple random experiments and associate probabilities with their outcomes.

Example 1

i) A balanced coin is tossed. The two possible outcomes are head (H) and tail (T). We associate probability $P\{H\} = 1/2$ to the outcome H and probability $P\{T\} = 1/2$ to T.

ii) A person is selected from a large group of persons and his blood group is determined. It can be one of the four blood groups O, A, B and AB. One possible assignment of probabilities to these outcomes is given below

Blood group	O	A	B	AB
Probability	0.34	0.27	0.31	0.08

Now look carefully at the probabilities attached to the sample points in Example 1 (i) and (ii). Did you notice that

i) these are numbers between 0 and 1, and

ii) the sum of the probabilities of all the sample points is one ?

This is not true of this example alone. In general, we have the following rule or axiom about the assignment of probabilities to the points of a discrete sample space.

Axiom : Let Ω be a discrete sample space containing the points $\omega_1, \omega_2, \ldots$; i.e.,

$$\Omega = \{\omega_1, \omega_2, \ldots\}.$$

To each point ω_j of Ω, assign a number $P\{\omega_j\}$, $0 \leq P\{\omega_j\} \leq 1$, such that

$$P\{\omega_1\} + P\{\omega_2\} + \ldots = 1. \qquad \ldots(1)$$

We call $P\{\omega_j\}$, the **probability of** ω_j.

Now see if you can do the following exercise on the basis of this axiom.

E1) A sample space Ω consists of eight points $\omega_1, \omega_2, \ldots, \omega_8$. Which of the following assignments of probabilities are valid ones ?

Assignment	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
(a)	1/8	1/8	1/8	1/8	1/4	0	0	1/4
(b)	1/4	0	0	1/4	0	0	0	0
(c)	1/16	2/16	3/16	4/16	5/16	6/16	7/16	–12/16
(d)	1/8	1/8	1/8	1/8	1/8	1/8	1/8	3/8
(e)	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

If you have done E1, you would have noticed that it is possible to have more than one valid assignment of probabilities to the same sample space. If the discrete sample space Ω is not finite, the left side of Equation (1) should be interpreted as an infinite series. For example, suppose $\Omega = \{\omega_1, \omega_2, \ldots\}$ and

$$P\{\omega_j\} = 1/2^j, \forall\ j = 1, 2, \ldots\ldots$$

If $|r| < 1$, the sum of the infinite geometric series $a + ar + ar^2 + \ldots$ is $\frac{a}{1-r}$.

Then this assignment is valid because, $0 \le P\{\omega_j\} \le 1$, and

$$P\{\omega_1\} + P\{\omega_2\} + \ldots = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \ldots$$

$$= \frac{1}{2}\left(1 + \frac{1}{2} + \frac{1}{2^2} + \ldots\ldots\right)$$

$$= 1.$$

So far we have not explained what the probability $P\{\omega_j\}$ assigned to the point ω_j signifies. We have just said that they are all arbitrary numbers between 0 and 1, except for the requirement that they add up to 1. In fact, we have not even tried to clarify the nature of the sample space except to assert that it be a discrete sample space. Such an approach is consistent with the usual procedure of beginning the study of a mathematical discipline with a few undefined notions and axioms and then building a theory based on the laws of logic (Remember the axioms of geometry ?). It is for this reason that this approach to the specification of probabilities to discrete sample spaces is called the **axiomatic approach.** It was introduced by the Russian mathematician A.N. Kolmogorov in 1933. This approach is mathematically precise and is now universally accepted. But when we try to use the mathematical theory of probability to solve some real life problems, that we have to interpret the significance of statements like "The probability of an event A is 0.6."

We now define the probability of an event A for a discrete sample space.

6.2.1 Probability of an Event : Definition

Let Ω be a discrete sample space consisting of the points $\omega_1, \omega_2, \ldots$, finite or infinite in number. Let $P\{\omega_1\}, P\{\omega_2\}, \ldots$ be the probabilities assigned to the points $\omega_1, \omega_2, \ldots$

P(A) is also called the probability of occurrence of A.

Definition 1 : The probability P(A) of an event A is the sum of the Probabilities of the points in A. More formally,

$$P(A) = \sum_{\omega_j \in A} P\{\omega_j\}, \ldots\ldots (2)$$

where $\sum_{\omega_j \in A}$ stands for the fact that the sum is taken over all the points $\omega_j \in A$. A is, of course, a subset of Ω. By convention, we assign probability zero to the empty set. Thus, $P(\Phi_j) = 0$.

The following example should help in clarifying this concept.

Example 2 : Let Ω be the sample space corresponding to three tosses of a coin with the following assignment of probabilities.

Sample point	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT

Probability	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

Let's find the probabilities of the events A and B, where

A : There is exactly one head in three tosses, and

B : All the three tosses yield the same result

Now A = {HTT, THT, TTH}

Therefore,

$$P(A) = 1/8 + 1/8 + 1/8 = 3/8.$$

Further, B = {HHH, TTT}. Therefore, $P(B) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$.

Proceeding along these lines you should be able to do this exercise.

E2) Let's denote the possible outcomes of Experiment 5 in Unit 1 as follows :

ω_1 = GGG, ω_2 = GGB, ω_3 = GBG, ω_4 = BGG,

ω_5 = BBG, ω_6 = BGB, ω_7 = GBB, ω_8 = BBB.

Consider the following assignment of probabilities.

$P\{\omega_1\} = (9/10)^3$, $P\{\omega_2\} = P\{\omega_3\} = P\{\omega_4\} = (9/10)^2\,(1/10)$

$P\{\omega_5\} = P\{\omega_6\} = P\{\omega_7\} = (9/10)\,(1/10)^2$, $P\{\omega_8\} = (1/10)^3$.

a) Verify that the above assignment of probabilities is valid.

b) Find the probability of getting

i) exactly one bad item (B)

ii) at least one good item (G).

A word about our notation and nomenclature is necessary at this stage. Although we say that $P\{\omega_j\}$ is the probability assigned to the point w_j of the sample space, it can be also interpreted as the probability of the singleton event $\{\omega_j\}$.

In fact, it would be useful to remember that probabilities are defined only for events and that $P\{\omega_j\}$ is the probability of the singleton event $\{\omega_j\}$. This type of distinction will be all the more necessary when you proceed to study probability theory for non-discrete sample spaces in Block 3.

Now let us look at some properties of the probabilities of events.

6.2.2 Probability of an Event : Properties

By now you know that the probability P(A) of an event A associated with a discrete sample space is the sum of the probabilities assigned to the sample points in A. In this section we discuss the properties of the probabilities of events.

P1 : For every event A, $0 \le P(A) \le 1$.

Proof : This is a straightforward consequence of the definition of P(A). Since it is the sum of non-negative numbers, $P(A) \ge 0$. Since the sum of the probabilities assigned to all the points in the sample space is one and since A is a subset of Ω, the sum of the probabilities assigned to the points in A cannot exceed $P(\Omega)$, which is one. In other words, whatever may be the event A, $0 \le P(A) \le 1$.

Now here is an important remark.

Remark 1 : If $A = \phi$, $P(\phi_j) = 0$. However, $P(A) = 0$ does not, in general, imply that A is the empty set. For example, consider the assignment (i) of E1. You must have already shown that it is valid. If $A = \{\omega_6, \omega_7\}$. $P(A) = 0$ but A is not empty.

Similarly P(Ω) = 1, does it follow that B = Ω ? No. Can you think of a counter example ? What about E1) i) again ? If we take B = $\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_8\}$, then P(B) = 1 but B ≠ Ω. In this connection, recall that the empty set ϕ and the whole space Ω were called the impossible event and the sure event, respectively. In future, an event A with probability P(A) = 0 will be called a **null** event and an event B of probability one, will be called an **almost sure** event.

This remark brings out the fact that the impossible event is a null event but that a null event is not the impossible event. Similarly, the sure event is an almost sure event but an almost sure event is not necessarily the sure event.

Let us take up another property now.

P2 : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof : Recall that according to the definition, $P(A \cup B)$ is the sum of the probabilities attached to the points of $A \cup B$, each point being considered only once. However, when we compute P(A) + P(B), a point in $A \cap B$ is included once in the computation of P(A) and once in the computation of P(B). Thus, the probabilities of points in $A \cap B$ get added twice in the computation of P(A) + P(B). If we subtract the probabilities of all points in $A \cap B$, from P(A) + P(B), then we shall be left with $P(A \cup B)$, i.e.,

$$P(A \cup B) = P(A) + P(B) - \sum_{\omega_j \in A \cap B} P\{\omega_j\}.$$

The last term in the above relation is, by definition, $P(A \cap B)$. Hence we have proved P2.

We now list some properties which follow from P1 and P2.

P3 : If A and B are disjoint events, then

$$P(A \cup B) = P(A) + P(B).$$

P4 : $P(A^c) = 1 - P(A)$.

P5 : $P(A \cup B) \leq P(A) + P(B)$

Why don't you try to prove these yourself? That's what we suggest in the following exercise.

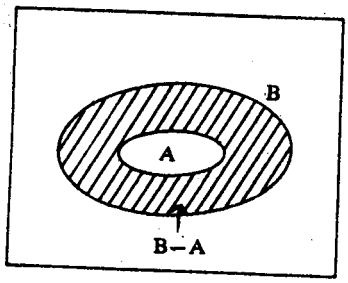


Fig. 1

E3) Prove P3, P4 and P5.

We continue with the list of properties.

P6 : If $A \subset B$, then $P(A) \leq P(B)$.

Proof : If $A \subset B$, A and B – A are disjoint events and their union, $A \cup (B - A)$ is B. Also see Fig. 1. Hence by P3,

$$P(B) = P(A \cup (B - A)) = P(A) + P(B - A).$$

Since by P1, $P(B - A) \geq 0$, P6 follows from the above equation.

Now let us take a look at P5 again.

The inequality $P(A \cup B) \leq P(A) + P(B)$ in P5 is sometimes called Boole's inequality. We claim that equality holds in Boole's inequality if $A \cap B$ is a null event. Do you agree?

An easy induction argument leads to the following generalisation of P5.

Boole's inequality : If $A_1, A_2, \ldots, A_N$ are N events, then

$$P\left(\bigcup_{j=1}^{N} A_j\right) \leq \sum_{j=1}^{N} P(A_j)$$

Proof : By P5, the result is true for N = 2. Assume that it is true for N ≤ r, and observe that $A_1 \cup A_2 \cup \ldots \cup A_{r+1}$ is the same as $B \cup A_{r+1}$, where $B = A_1 \cup A_2 \cup \ldots \cup A_r$. Then by P5,

$$P\left(\bigcup_{j=1}^{r+1} A_j\right) = P(B \cup A_{r+1}) \leq P(B) + P(A_{r+1})$$

$$\leq \sum_{j=1}^{r} P(A_j) + P(A_{r+1}),$$

where the last inequality is a consequence of the induction hypothesis. Hence, if Boole's inequality holds for $N \leq r$, it holds for $N = r + 1$ and hence for all $N \geq 2$.

A similar induction argument yields

P7 : If $A_1, A_2, \ldots, A_n$ are pair wise disjoint events, i.e., if $A_i \cap A_j = \phi$, $i \neq j$, then

$$P\left(\bigcup_{j=1}^{n} A_j\right) = P(A_1) + P(A_2) + \ldots + P(A_n). \qquad \ldots (3)$$

We sometimes refer to the relation (3) as the **Property of finite additivity.**

We can generalise P7 to apply to an infinite sequence of events.

P8 : If $\{A_n, n \geq 1\}$ is a sequence of pair wise disjoint events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j) \qquad \ldots (4)$$

P8 is called the **σ-additivity property.**

In the general theory of probability, which covers non-discrete sample spaces as well, σ-additivity and therefore finite additivity is included as an axiom to be satisfied by probabilities of events.

We now discuss some examples based on the above properties.

Example 3 : Let us check whether the probabilities P(A) and P(B) are consistently defined in the following cases.

i) $P(A) = 0.3\ P(B) = 0.4, P(A \cap B) = 0.4$

ii) $P(A) = 0.3\ P(B) = 0.4, P(A \cap B) = 0.8$

Here we have to see whether P1, P2, P3, P5 and P6 are satisfied or not. P4, P7 and P8 do not apply here since we are considering only two sets. In both the cases P(A) and P(B) are not consistently defined. Since $A \cap B \subset A$, by P6, $P(A \cap B) \leq P(A)$. In case (i), $P(A \cap B) = 0.4 > 0.3 = P(A)$, which is impossible. Similar is the situation with case (ii). Moreover, note that case (ii) also violates P1 and P2. Recall that by P2,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

but $P(A) + P(B) - P(A \cap B) = 0.3 + 0.4 - 0.8 = -0.1$ which is impossible.

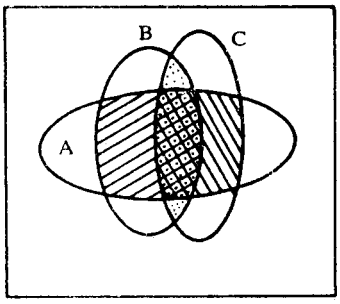


Fig. 2

Example 4 : We can extend the property P2 to the case of three events, i.e., we can show that

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B)$$
$$- P(B \cap C) - P(C \cap A) + P(A \cap B \cap C) \qquad \ldots (5)$$

Denote $B \cup C$ by H. Then $A \cup B \cup C = A \cup H$ and by P2, $P(A \cup B \cup C)$

$$= P(A \cup H) = P(A) + P(H) - P(A \cap H.) \qquad \ldots (6)$$

But $P(H) = P(B \cup C) = P(B) + P(C) - P(B \cap C)$ $\ldots (7)$

and $P(A \cap H) = P(A \cap (B \cup C))$

$$= P((A \cap B) \cup (A \cap C))$$
$$= P(A \cap B) + P(A \cap C) - P\{(A \cap B) \cap (A \cap C)\}$$
$$= P(A \cap B) + P(A \cap C) - P(A \cap B \cap C) \qquad \ldots (8)$$

Substituting from (7) and (8) in (6) we get the required result. Also see Fig. 2.

Here are some simple exercises which you can solve by using P1-P7

E4) Prove the following :

a) If $P(A) = P(B) = 1$, then $P(A \cup B) = P(A \cap B) = 1$.

b) If $P(A) = P(B) = P(C) = 0$, then $P(A \cup B \cup C) = 0$.

c) We have mentioned that by convention we take $P(\phi) = 0$.

But see if you can prove it by using P4.

E5) Fill in the blanks in the following table :

P(A)	P(B)	P(A ∪ B)	P(A ∩ B)
0.4	0.8		0.3
	0.5	0.6	0.25

E6) Explain why each one of the following statements is incorrect.

a) The probability that a student will pass an examination is 0.65 and that he would fail is 0.45.

b) The probability that team A would win a match is 0.75, that the game will end in a draw is 0.15 and that team A will not loose the game is 0.95.

c) The following is the table of probabilities for printing mistakes in a book.

No. of printing mistakes	0	1	2	3	4	5 or more
Probability	0.12	0.25	0.36	0.14	0.09	0.07

d) The probabilities that a bank will get 0, 1, 2, or more than 2 bad cheques on a given day are 0.08, 0.21, 0.29 and 0.40, respectively.

E7) There are two assistants Seema (S) and Wilson (W) in an office. The probability that Seema will be absent on any given day is 0.05 and that Wilson will be absent on any given day is 0.10. The probability that both will be absent on the same day is 0.02. Find the probability that on a given day,

a) both Seema and Wilson would be present,

b) at least one of them would be present, and

c) only one of them will be absent.

E8) A large office has three xerox machines M_1, M_2 and M_3. The probability that on a given day

M_1 works is 0.60

M_2 works is 0.75

M_3 works is 0.80

both M_1 and M_2 work is 0.50

both M_1 and M_3 work is 0.40

both M_2 and M_3 work is 0.70

all of them work is 0.25.

Find the probability that on a given day at least one of the three machines works.

Through the examples and exercises in this section we hope you have grasped the axiomatic approach to probability. In the next section we'll describe the classical approach.

6.3 CLASSICAL DEFINITION OF PROBABILITY

In the early stages, probability theory was mainly concerned with its applications to games of chance. The sample space for these games consisted of a finite number of outcomes. These simple situations led to a definition of probability which is now called the classical definition. It has many limitations. For example, it cannot be applied to infinite sample

space. However, it is useful in understanding the concept of randomness so essential in the planning of experiments, small and large-scale sample surveys, as well as in solving some interesting problems. We shall motivate the classical definition with some examples. We shall then formulate the classical definition and apply it to solve some simple problems.

Suppose we toss a coin. This experiment has only two possible outcomes : Head (H) and Tail (T). If the coin is a balanced coin and is symmetric, there is no particular reason to expect that H is more likely than T or that T is more likely than H. In other words, we may assume that the two outcomes H and T have the same probability or that they are equally likely. If they have the same probability, and if the sum of the two probabilities P{H} and P{T} is to be one, we must have $P\{H\} = P\{T\} = 1/2$.

Similarly, if we roll a symmetric, balanced die once, we should assign the same probability, viz. 1/6 to each of the six possible outcomes $1, 2, \ldots, 6$.

The same type of argument, when used for assigning probabilities to the results of drawing a card from a well-shuffled pack of 52 playing cards leads us to say that the probability of drawing any specified card is 1/52.

In general, we have the following :

Definition 2 : Suppose a sample space Ω has a finite number n of points $\omega_1, \omega_2, \ldots, \omega_n$. The classical definition assigns the probability 1/n to each of these points, i.e.,

$$P\{\omega_j\} = \frac{1}{n}, j = 1, \ldots, n.$$

The above assignment is also referred to as the assignment in case of equally likely outcomes. You can check that in this case, the total of the probabilities of all the n points is $n \times \frac{1}{n} = 1$. In fact, this is a valid assignment even from the axiomatic point of view.

Now suppose that an event A contains m points. Then under the classical assignment, the probability P(A) of A is m/n. The early probabilists called m, the number of cases favourable to A and n, the total number of cases. Thus, according to the classical definition,

$$P(A) = \frac{\text{Number of cases favourable to A}}{\text{Total number of cases}}$$

We have already mentioned that this is a valid assignment consistent with the Axiom in Sec. 6.2. Therefore, it follows that the probabilities of events, defined in this manner, possess the properties $P_1 - P_7$.

We now give some examples based on this definition.

Example 5 : Two identical symmetric dice are thrown. Let us find the probability of obtaining a total score of 8.

The total number of possible outcomes is $6 \times 6 = 36$. There are 5 sample points, (2, 6), (3, 5), (4, 4), (5, 3), (6, 2), which are favourable to the event A of getting a total score of 8. Hence the required probability is 5/36.

Example 6 : If each card of an ordinary deck of 52 playing cards has the same probability of being drawn, let us find the probability of drawing.

i) a red king or a black ace

ii) a 3, 4, 5, 6 or 8 ?

Let's tackle these one by one

i) Since there are two red kings (diamond and heart) and two black aces (spade and club), the number of favourable cases is 4. The required probability is $4/52 = 1/13$.

ii) There are 4 cards of each of the 5 denominations 3, 4, 5, 6 and 8. Thus, the total number of favourable cases is 20 and the required probability is $20/52 = 5/13$.

You must have realised by this time that in order to apply the classical definition of probability, you must be able to count the number of points favourable to an event A as well as the total number of sample points. This is not always easy. We can, however, use the theory of permutations and combinations for this purpose.

To refresh your memory, here we give two important rules which are used in counting.

1) **Multiplication Rule :** If an operation is performed in n_1 ways and for each of these n_1 ways, a second operation can be performed in n_2 ways, then the two operations can be performed together in $n_1 n_2$ ways. See Fig. 3.

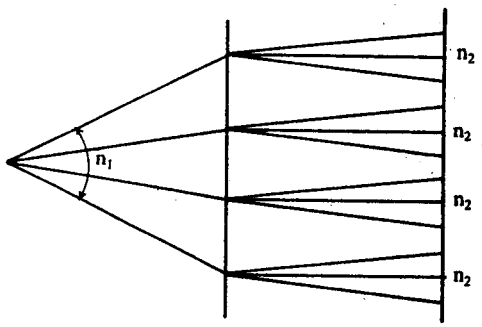


Fig. 3

2) **Addition Rule :** Suppose an operation can be performed in n_1 ways and a second operation can be performed in n_2 ways. Suppose, further that it is not possible to perform both together. Then the number of ways in which we can perform the first or the second operation in $n_1 + n_2$. See Fig. 4.

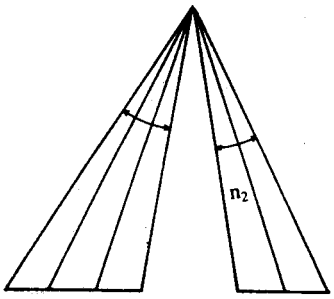


Fig. 4

We now illustrate the use of this theory in calculating probabilities by considering some examples. We **assume** that all outcomes in each of these examples are equally likely. Under this assumption, the classical definition of probability is applicable.

Example 7 : We first select a digit out of the ten digits, 0, 1, 2, 3, . . ., 9. Then we select another digit out of the remaining nine. What will be the probability that both these digits are odd?

We can select the first digit in 10 ways and for each of these ways we can select the second digit in 9 ways. Therefore, the total number of points in the sample space is $10 \times 9 = 90$. The first digit, can be odd in 5 ways (1, 3, 5, 7, 9), and then the second digit can be odd in 4 ways. Thus, the total number of ways in which both the digits can be odd is $5 \times 4 = 20$. The required probability is therefore $\frac{20}{90} = \frac{2}{9}$.

This is called selection **without replacement** since we do not replace the first selected digit bac before the second selection.

Remark 2 : In Example 7, every digit had the same chance of being selected. This is sometimes expressed by saying that the digits were selected **at random** (with equal probability). Selection at random is generally taken to be synonymous with the assignment of the same probability to all the sample points, unless stated otherwise.

We now give a number of examples to show how to calculate the probabilities of events in a variety of situations. Please go through these examples carefully. If you understand them, you will have no difficulty in doing the exercises later.

Example 8 : A box contains ninety good and ten defective screws. Let us find the probability that 5 screws selected at random out of this box are all good.

Let A be the event that the 5 selected screws are all good.

Now we can choose 5 screws out of 100 screws in $\binom{100}{5}$ ways. If the selected 5 screws are to be good, they will have to be selected out of the 90 good screws. This can be done in $\binom{90}{5}$ ways. This is the number of sample points favourable to A. Hence the probability of A

$$\frac{\binom{90}{5}}{\binom{100}{5}} = \frac{90 \times 89 \times 88 \times 87 \times 86}{100 \times 99 \times 98 \times 97 \times 96} \approx 0.58.$$

Example 9 : A government prints 10 lakh lottery tickets of value of Rs. 2 each. We would like to know the number of tickets that must be bought to have a chance of 0.5 or more to win the first prize of 2 lakhs.

The prize-winning ticket can be randomly selected out of the 10 lakh tickets in 10^6 ways.

Now, let m denote the number of tickets that we must buy. Then m is the number of points favourable to our winning the first prize. Therefore, the probability of our winning the first prize, is, $\frac{m}{10^6}$.

Since we want that $\frac{m}{10^6} \geq \frac{1}{2}$, therefore $m \geq \frac{10^6}{2}$. This means that we must buy at least $\frac{10^6}{2} = 500,000$ tickets, at a cost of at least Rs. 10 lakhs ! Not a profitable proposition at all !

Example 10 : In a study centre batch of 100 students, 54 opted for MTE-06, 69 opted for MTE-11 and 35 opted for both MTE-06 and MTE-11. If one of these students is selected at random, let us find the probability that the student has opted for MTE-06 or MTE-11.

Let M denote the event that the randomly selected student has opted for MTE-06 and S the event that he/she has opted for MTE- 11. We want to know $P(M \cup S)$. According to the classical definition, $P(M) = \frac{54}{100}$, $P(S) = \frac{69}{100}$ and $P(M \cap S) = \frac{35}{100}$. Thus

$$P(M \cup S) = P(M) + P(S) - P(M \cap S)$$

$$\frac{54}{100} + \frac{69}{100} - \frac{35}{100} = \frac{88}{100} = 0.88.$$

Suppose now we want to know the probability that the randomly selected student has opted for neither MTE-06 nor MTE-11. This means that we want to know $P[M^c \cap S^c]$.

Now,

$$M^c \cap S^c = (M \cup S)^c$$

Recall De Morgan's laws from Unit 1, MTE-04

Therefore,

$$P(M^c \cap S^c) = 1 - P[M \cup S] = 1 - 0.88 = 0.12.$$

The adjective 'unbiased' attached to dice implies that all the sample points are equiprobable, i.e., have equal probabilities of occurrence.

Lastly, to obtain the probability that the student has opted for MTE-06 but not for MTE-11, i.e., to obtain $P(M \cap S^c)$, observe that $M = (M \cap S) \cup (M \cap S^c)$ and that $M \cap S$ and $M \cap S^c$ are disjoint events. Thus.

$$P(M) = P(M \cap S) + P(M \cap S^c)$$

$$\text{or } P(M \cap S^c) = P(M) - P(M \cap S)$$

$$= \frac{54}{100} - \frac{35}{100} = 0.19.$$

Example 11 : A throws six unbiased dice and wins if he has at least one six. B throws twelve unbiased dice and wins if he has at least two sixes. Who do you think is more likely to win?

We would urge you to make a guess first and then go through the following computations.

Check if your intuition was correct.

The total number of outcomes for A is $n_A = 6^6$ and that for B is $n_B = 6^{12}$. We will first

calculate the probabilities q_A and q_B that A and B, respectively, loose their games. Then the probabilities of their winning are $P_A = 1 - q_A$ and $P_B = 1 - q_B$, respectively. We do this because q_A and q_B are easier to compute.

Now A loses if he does not have a six on any of the 6 dice he rolls. This can happen in 5^6 different ways, since he can have no six on each die in 5 ways. Hence $q_A = 5^6/6^6$ and therefore, $P_A = 1 - (5/6)^6 \cong 0.665$.

In order to calculate q_B, observe that B loses if he has no six or exactly one six. The probability that he has no six is $5^{12}/6^{12} = (5/6)^{12}$. Now the single six can occur on any one of the 12 dice, i.e., in $\binom{12}{1}$ ways. Then all the remaining 11 dice have to have a score other than six. This can happen in 5^{11} ways.

Therefore, the total number of ways of obtaining one six is $\binom{12}{1}5^{11}$. Hence the probability that B has exactly one six is $\frac{12 \times 5^{11}}{6^{12}}$.

The events of "no six" and "one six" in the throwing of 12 dice are disjoint events. Hence the probability

$$q_B = (5/6)^{12} + 12\,\frac{5^{11}}{6^{12}} \cong 0.381$$

Thus, $P_B \cong 1 - 0.381 = 0.619$.

Comparing P_A and P_B, we can conclude that A has a greater probability of winning.

Now here are some exercises which you should try to solve.

E9) Two cards are drawn in succession from a deck of 52 playing cards with replacement. What is the probability that both cards are of denomination greater than 2 and less than 5?

E10) If 3 books are selected at random from a shelf containing 5 novels, 3 books of poems and a dictionary, what is the probability that

a) dictionary is not selected

b) 2 novels and 1 book of poems are selected.

E11) A person has 4 keys of which only one fits the lock. He tries them successively at random without replacement. This procedure may require 1, 2, 3 or 4 attempts. Show that the probability of any one of these 4 outcomes is 1/4.

E12) In an experiment to study the dependence of blood pressure on smoking habits, the following data were collected on 220 individuals.

	Non-smoker	Moderate smoker	Heavy smoker
High blood pressure	20	40	40
Normal blood pressure	60	30	30

One of the persons is selected at random. What is the probability that he is

a) a smoker with high blood pressure

b) a non-smoker with normal blood pressure

c) a smoker.

E13) Two balanced dice are thrown. What is the probability that the total score exceeds 8?

So far we have seen various examples of assigning probabilities to sample points and have

also discussed some properties of probabilities of events. In the next section we shall talk about the concept of conditional probability.

6.4 CONDITIONAL PROBABILITY

Suppose that two series of tickets are issued for a lottery. Let 1, 2, 3, 4, 5 be the numbers on the 5 tickets in series I and let 6, 7, 8, 9, be the numbers on the 4 tickets in series II. I hold the ticket bearing number 3. Suppose the first prize in the lottery is decided by selecting one of the 5 + 4 = 9 numbers at random. The probability that I will win the prize is 1/9. Does this probability change if it is known that the prize-winning ticket is from series I? In effect, we want to know the probability of my winning the prize, conditional on the knowledge that the prize-winning ticket is from series I.

In order to answer this question, observe that the given information reduces our sample-space from the set {1, 2, 3, 4, 5, 6, 7, 8, 9} to its subset {1, 2, 3, 4, 5} containing 5 points. In fact, this subset {1, 2, 3, 4, 5} corresponds to the event H that the prize winning ticket belongs to series I. If the prize winning ticket is selected by choosing one of these 5 numbers at random, the probability that I will win the prize is 1/5. Therefore, it seems logical to say that the conditional probability of the event A of my winning the prize, given that the prize-winning number is from series I, is

$$P(A \mid H) = 1/5.$$

Here $P(A \mid H)$ is read as the conditional probability of A given the event H. Note that we can write

$$P(A \mid H) = \frac{1/9}{5/9} = \frac{P(A \cap H)}{P(H)}$$

This discussion enables us to introduce the following formal definition. In what follows we assume that we are given a random experiment with discrete sample space Ω, and all relevant events are subsets of Ω.

Definition 3 : Let H be an event of positive probability, that is, $P(H) > 0$. The **conditional probability** $P(A \mid H)$ of an event A, given the event H, is

$$P(A \mid H) = \frac{P(A \cap H)}{P(H)} \qquad \ldots (9)$$

Notice that we have not put any restriction on the event A except that A and H be subsets of the same sample space Ω and that $P(H) > 0$.

Now we give two examples to help clarify this concept.

Example 12 : In a small town of 1000 people there are 400 females and 200 colour-blind persons. It is known that ten per cent, i.e. 40, of the 400 females are colour-blind. Let us find the probability that a randomly chosen person is colour-blind, given that the selected person is a female.

Now suppose we denote by A the event that the randomly chosen person is colour-blind and by H the event that the randomly chosen person is a female. You can see that

$$P(A \cap H) = 40/1000 = 0.04 \text{ and that}$$

$$P(H) = 400/1000 = 0.4.$$

Then

$$P(A \mid H) = \frac{P(A \cap H)}{P(H)} = \frac{0.04}{0.40} = 0.1.$$

Now can you find the probability that a randomly chosen person is colour-blind, given that the selected person is a male ?

If you denote by M the event that the selected person is a male, then

$$P(M) = \frac{600}{1000} = 0.6 \text{ and}$$

$$P(A \cap M) = \frac{160}{1000} = 0.16.$$

Therefore, $P(A \mid M) = \frac{\quad}{0.6} \approx 0.266$.

You must have noticed that $P(A \mid M) > P(A \mid H)$. So there are greater chances of a man being colour-blind as compared to a woman.

Example 13 : A manufacturer of automobile parts knows from past experience that the probability that an order will be completed on time is 0.75. The probability that an order is completed and delivered on time is 0.60. Can you help him to find the probability that an order will be delivered on time given that it is completed ?

Let A be the event that an order is delivered on time and H the event that it is completed on time. Then $P(H) = 0.75$ and $P(A \cap H) = 0.60$. We need $P(A \mid H)$.

$$P(A \mid H) = \frac{P(A \cap H)}{P(H)} = \frac{0.60}{0.75} = 0.8.$$

Have you understood the definition of conditional probability? You can find out for yourself by doing these simple exercises.

E14) If A is the event that a person suffers from high blood pressure and B is the event that he is a smoker, explain in words what the following probabilities represent.

a) $P(A \mid B)$

b) $P(A^c \mid B)$

c) $P(A \mid B^c)$

d) $P(A^c \mid B^c)$.

E15) Two unbiased dice are rolled. They both show the same score. What is the probability that their common score is 6?

We now state some of the properties of $P(A \mid H)$.

P'1 : For any set A, $0 \leq P(A \mid H) \leq 1$.

Recall that since $A \cap H \subset H$, $P(A \cap H) \leq P(H)$. The required property follows immediately.

P'2 : $P(A \mid H) = 0$ if and only if $A \cap H$ is a null set. In particular, $P(\phi \mid H) = 0$ and $P(A \mid H) = 0$ if A and H are disjoint events.

P'3 : $P(A \mid H) = 1$ if and only if $P(A \cap H) = P(H)$.

In particular,

$$P(\Omega \mid H) = 1 \text{ and } P(H \mid H) = 1$$

P'4 : $P(A \cup B \mid H) = P(A \mid H) + P(B \mid H) - P(A \cap B \mid H)$.

How do we get P'4 ? Well, since

$$(A \cup B) \cap H = (A \cap H) \cup (B \cap H),$$

P_2 gives us

$$P((A \cup B) \cap H) = P(A \cap H) + P(B \cap H) - P(A \cap B \cap H).$$

Now use the definition of the conditional probability to obtain P'4.

Using P'4 and P3 and P4 of Sec. 6.2.2, we get

P'5 : If A and B are disjoint events, $P(A \cup B \mid H) = P(A \mid H) + P(B \mid H)$

and $P(A^c \mid H) = 1 - P(A \mid H)$

Compare $P'_1 - P'_5$ with the properties of (unconditional) probabilities given in Sec. 6.2.2. You will find that the conditional probabilities, given the event H, have all the properties of unconditional probabilities, which are sometimes called the absolute properties.

We can use the conditional probabilities to compute the unconditional probabilities of events by employing the following obvious fact,

$$P(A \cap H) = P(H)\, P(A \mid H). \qquad \ldots (10)$$

obtained from Definition 3 of $P(A \mid H)$.

See 1.4 for the interpretations of $P(A \cap H)$ and $P(A \mid H)$.

Here is an important remark related to (10).

Remark 3 : Relation (10) holds even if $P(H) = 0$, provided we interpret $P(A \mid H) = 0$ if $P(H) = 0$. In words, this means that if the probability of occurrence of H is zero, we say that the probability of occurrence of A, given that H has occurred, is also zero. This is so, because $P(H) = 0$ implies $P(A \cap H) = 0$, $(A \cap H)$ being a subset of H.

We now give an example to illustrate the use of Relation (10).

Example 14 : Two cards are drawn at random and without replacement from a pack of 52 playing cards. Let us find the probability that both the cards are red.

Let A_1 and A_2 denote, respectively the events that cards drawn on the first and second draw are red. Then by the classical definition, $P(A_1) = 26/52$, since there are 26 red cards. If the first card is red, we are left with 25 red cards in the pack of 51 cards. Hence $P(A_2 \mid A_1) = 25/51$. Thus, the probability $P(A_1 \cap A_2)$ of both cards being red is

$$P(A_1 \cap A_2) = P(A_1)\, P(A_2 \mid A_1)$$

$$= \frac{26}{52} \cdot \frac{25}{51} \approx 0.245.$$

Relation (10) specifies the probability of $A \cap H$ in terms of $P(H)$ and $P(A/H)$. We can extend this relation to obtain the probability, $P(A_1 \cap A_2 \cap A_3)$ in terms of $P(A_1)$, $P(A_2 \mid A_1)$ and $P(A_3 \mid A_1 \cap A_2)$. We, of course, assume that $P(A_1)$ and $P(A_1 \cap A_2)$ are both positive. Can you guess what this relation could be? Suppose we write

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)\, \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)}$$

Does this give you any clue? This gives us,

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2).$$

Now let us use this to compute some probabilities.

Example 15 : A box of mangoes is inspected by examining three randomly selected mangoes drawn without replacement. If all the three mangoes are good, the box is sent to the market, otherwise it is rejected. Let us calculate the probability that a box of 100 mangoes containing 90 good mangoes and 10 bad ones will pass the inspection.

Let A_1, A_2 and A_3, respectively denote the events that the first, second and third mangoes are good. Then $P(A_1) = 90/100$, $P(A_2 \mid A_1) = 89/99$, and $P(A_3 \mid A_1 \cap A_2) = 88/98$ according to the classical definition. Thus,

$$P(A_1 \cap A_2 \cap A_3) = \frac{90}{100} \cdot \frac{89}{99} \cdot \frac{88}{98} \approx 0.727.$$

We end this section with a derivation of a well-known theorem in probability theory, called the Bayes' theorem.

Consider an event B and its complementary event B^c. The pair (B, B^c) is called a partition of Ω, since they satisfy $B \cap B^c = \phi$, and $B \cup B^c$ is the whole sample space Ω. Observe that for any event A,

$$A = A \cap \Omega = A \cap (B \cup B^c) = (A \cap B) \cup (A \cap B^c).$$

Since $A \cap B$ and $A \cap B^c$ are subsets of the disjoint sets B and B^c, respectively, they themselves are disjoint. As a consequence, $P(A) = P(A \cap B) + P(A \cap B^c)$.

Now using Relation (10), we have

$$P(A) = P(B)\, P(A \mid B) + P(B^c)\, P(A \mid B^c). \qquad \ldots (11)$$

Here we do not insist that $P(B)$ and $P(B^c)$ be positive and follow the convention stated in Remark 3.

It is now possible to extend Equation (11) to the case when we have a partition of Ω consisting of more than two sets. More specifically, we say that the n sets $B_1, B_2, \ldots, B_n$ constitute a partition of Ω if any two of them are disjoint, i.e.,

$$B_i \cap B_j = \phi, i \neq j, i, j = 1, \ldots, n$$

and their union is Ω, i.e.,

$$\bigcup_{j=1}^{n} B_j = \Omega.$$

We can now write for any event A,

$$A = A \cap \Omega = A \cap \left(\bigcup_{j=1}^{n} B_j \right) = \bigcup_{j=1}^{n} (A \cap B_j).$$

Since $A \cap B_i$ and $A \cap B_j$ are respectively subsets of B_i and B_j, $i \neq j$, they are disjoint. Consequently by P7,

$$P(A) = \sum_{j=1}^{n} P(A \cap B_j)$$

$$\text{or } P(A) = \sum_{j=1}^{n} P(B_j)\, P(A \mid B_j), \qquad \ldots (12)$$

which is obtained by using (10). This result (12) leads to the celebrated Bayes' theorem, which we now state.

Theorem 1 (Bayes' Theorem) : If $B_1, B_2, \ldots, B_n$ are n events which constitute a partition of Ω and A is an event of positive probability, then

$$P(B_r \mid A) = \frac{P(B_r)\, P(A \mid B_r)}{\sum_{j=1}^{n} P(B_j)\, P(A \mid B_j)}$$

for any r, $1 \leq r \leq n$.

Proof : Observe that by definition,

$$P(B_r \mid A) = \frac{P(A \cap B_r)}{P(A)}$$

$$= \frac{P(B_r)\, P(A \mid B_r)}{P(A)}, \qquad \text{by (10)}$$

$$= \frac{P(B_r)\, P(A \mid B_r)}{\sum_{j=1}^{n} P(B_j)\, P(A \mid B_j)}, \qquad \text{by (12).}$$

The proof is complete.

In the examples that follow, you will see a variety of situations in which Bayes' theorem is useful.

Example 16 : It is known that 25 per cent of the people in a community suffer from TB. A test to diagnose this disease is such that the probability is 0.99 that a person suffering from it will show a positive result indicating its presence. The same test has probability 0.20 that a person not suffering from TB has a positive test result. If a randomly selected person from the community has positive test result, let us find the probability that he has TB.

Let B_1 denote the event that a randomly selected person has TB. Let $B_2 = B_1^c$. Then from the given information, $P(B_1) = 0.25$, $P(B_2) = 0.75$. Let A denote the event that the test for the

randomly selected person yields a positive result. Then $P(A \mid B_1) = 0.99$ and $P(A \mid B_2) = 0.20$. We need to obtain $P(B_1 \mid A)$. By applying Bayes' theorem we get

$$P(B_1 \mid A) = \frac{P(B_1)\,P(A \mid B_1)}{P(B_1)\,P(A \mid B_1) + P(B_2)\,P(A \mid B_2)}$$

$$= \frac{0.25 \times 0.99}{0.25 \times 0.99 + 0.75 \times 0.20}$$

$$\approx 0.623.$$

Example 17 : We have three boxes, each containing two covered compartments. The first box has a gold coin in each compartment. The second box has a gold coin in one compartment and a silver coin in the other. The third box has a silver coin in each of its compartments. We choose a box at random and open a drawer at random. It contains a gold coin. We would like to know the probability that the other compartment also has a gold coin.

Let B_1, B_2, B_3, respectively, denote the events that Box 1, Box 2 and Box 3 are selected. It is easy to see that B_1, B_2, B_3 constitute a partition of the sample space of the experiment.

Since the boxes are selected at random, we have

$$P(B_1) = P(B_2) = P(B_3) = 1/3.$$

Let A denote the event that a gold coin is located. The composition of the boxes implies that

$$P(A \mid B_1) = 1,\ P(A \mid B_2) = 1/2,\ P(A \mid B_3) = 0.$$

Since one gold coin is observed, we will have a gold coin in the other unobserved compartment of the box only if we have selected Box 1. Thus, we need to obtain $P(B_1 \mid A)$. Now by Bayes Theorem

$$P(B_1 \mid A) = \frac{P(B_1)\,P(A \mid B_1)}{P(B_1)\,P(A \mid B_1) + P(B_2)\,P(A \mid B_2) + P(B_3)\,P(A \mid B_3)}$$

$$= \frac{(1/3) \times 1}{(1/3) \times 1 + (1/3) \times 1/2 + (1/3) \times 0}$$

$$= 2/3.$$

Do you feel confident enough to try and solve these exercises now? In each of them, the crucial step is to define the relevant events properly. Once you do that, the actual calculation of probabilities is child's play.

E16) In a city the weather changes frequently. It is known from past experience that a rainy day is followed by a sunny day with probability 0.4, and that a sunny day is followed by a rainy day with probability 0.7. Assume that the weather on any given day depends only on the weather of the previous day. Find the probability that

a) a rainy day is followed by a rainy day

b) it would rain on Saturday and Sunday when Friday was rainy

c) the entire period from Monday to Friday is rainy given, that the previous Sunday was sunny.

This is an example of a **Markov chain**, named after the Russian mathematician A. Markov (1856–1922) who initiated their study

E17) An urn contains 4 white and 4 black balls. A ball is drawn at random, its colour is noted and is returned to the urn. Moreover, 2 additional balls of the colour drawn are put in the urn and then a ball is drawn at random. What is the probability that the second ball is black?

This procedure is called **Polya's urn scheme.**

E18) In a community 2 per cent of the people suffer from cancer. The probability that a doctor is able to correctly diagnose a person with cancer as suffering from cancer is 0.80. The doctor wrongly diagnoses a person without cancer as having cancer with probability 0.05. What is the probability that a randomly selected person diagnosed as having cancer is really suffering from cancer?

E19) An explosion in a factory manufacturing explosives can occur because of (i) leakage of electricity, (ii) defects in machinery, (iii) carelessness of workers or (iv) sabotage. The probability that

i) there is a leakage of electricity is 0.20

ii) the machinery is defective is 0.30

iii) the workers are careless is 0.40

iv) there is sabotage is 0.10

The engineers feel that an explosion can occur with probability (i) 0.25 because of leakage of electricity, (ii) 0.20 because of defects in the machinery, (iii) 0.50 because of carelessness of workers, and (iv) 0.75 because of sabotage. Which is the most likely cause of explosion?

Using the concept of conditional probability, we now introduce independent events in the next section.

6.5 INDEPENDENCE OF EVENTS

From the examples discussed in the previous section you know that the conditional probability $P(A \mid H)$ is, in general, not the same as the unconditional probability $P(A)$. Thus, the knowledge of H affects the chances of occurrence of A. The following example illustrates this fact more explicitly.

Example 18 : A box has 4 tickets numbered 1, 2, 3 and 4. One of these tickets is drawn at random. Let $A = \{1, 2\}$ be the event that the randomly selected ticket bears the number 1 or 2. Similarly define $B = \{1\}$. Then

$$P(A) = 1/2,\ P(B) = 1/4;\ \text{and}\ P(A \cap B) = 1/4.$$

Therefore, $P(B \mid A) = (1/4) / (1/2) = 1/2$.

So we have $\mathbf{P(B \mid A) > P(B).}$

On the other hand, if $C = \{1, 2, 3\}$ and $D = \{1, 2, 4\}$, then $P(C) = P(D) = 3/4$ and $P(C \cap D) = 1/2$. Thus,

$$P(D \mid C) = \frac{1/2}{3/4} = 2/3, \text{ and in this case,}$$

$$\mathbf{P(D \mid C) < P(D).}$$

This example illustrates that additional information (about the occurrence of an event) can increase or decrease the probability of occurrence of another event. We would be interested in those situations which correspond to the cases when $P(B \mid A) = P(B)$, as in the following example.

Example 19 : We continue with the previous example. But now define $H = \{1, 2\}$ and $K = \{1, 3\}$. Then

$$P(H) = 1/2,\ P(K) = 1/2 \text{ and } P(H \cap K) = 1/4.$$

Hence

$$P(K \mid H) = \frac{1/4}{1/2} = 1/2 = P(K).$$

In this example, knowledge of the occurrence of H does not alter the probability of occurrence of K. We call such events, independent events.

Thus, two events A and B are independent, if

$$P(B \mid A) = P(B). \qquad \ldots (13)$$

However, in this definition, we need to have $P(A) > 0$. Using the definition of $P(B \mid A)$, we can rewrite (13) as

$$P(A \cap B) = P(A)\,P(B), \qquad \ldots (14)$$

which does not require that $P(A)$ or $P(B)$ be positive. We shall now use (14) to define independence of two events.

Definition 4 : Let A and B be two events associated with the same random experiment. They are said to be **stochastically independent** or simply **independent** if

$$P(A \cap B) = P(A)\,P(B).$$

So the events A and B in Example 18 are not independent. Similarly, events C and D are also not independent. But events K and H in Example 19 are independent.

See if you can apply Definition 4 and solve this exercise.

E20) Two unbiased dice are rolled. Let

A_1 be the event "odd face with the first die"

A_2 be the event "odd face with the second die"

B_1 be the event that the score on the first die is 1

B_2 be the event that the total score is at most 3.

Check the independence of the events

a) A_1 and A_2

b) B_1 and B_2

We now proceed to study some implications of independence of two events A_1 and A_2. Recall that

$$P(A_1) = P(A_1 \cap A_2) + P(A_1 \cap A_2^c).$$

Then

$$P(A_1 \cap A_2^c) = P(A_1) - P(A_1 \cap A_2)$$

Now, if A_1 and A_2 are independent, we get

$$P(A_1 \cap A_2^c) = P(A_1)\{1 - P(A_2)\}$$

$$= P(A_1)\, P(A_2^c).$$

Thus, the independence of A_1 andA_2 implies that of A_1 and A_2^c. Now interchange the roles of A_1 and A_2. What do you get? We get that if A_1 and A_2 are independent, then so are A_1^c and A_2. The independence of A_1^c and A_2 then implies the independence of A_1^c and A_2^c too.

Now here is an interesting fact.

If A is an almost sure event, then A and another event B are independent.

Let us see how. Since A is an almost sure event, P(A) = 1. Hence $P(A^c) = 0$ and therefore, $P(A^c \cap B) = 0$. In particular,

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(A \cap B).$$

One consequence of this is that

$$P(A \cap B) = 1.P(B) = P(A)\, P(B),$$

which implies that A and B are independent.

Can you prove a similar result for a null event ? You can check that if A is a null event, then A and any other event B are independent.

Now, can we extend the definition of independence of two events to that of the independence of three events? The obvious way seems to be to call A_1, A_2, A_3 independent if $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$. **But this does not work.** Because if 3 events are independent, we would expect any two of them also to be independent. But this is not ensured by the condition above. To appreciate this, consider the case when $A_1 = A_2 = A$, $0 < P(A) < 1$, and $P(A_3) = 0$. Then $P(A_1 \cap A_2) = P(A) \neq P(A_1)\, P(A_2) = P[(A)]^2$.

Thus, A_1 and A_2 are not independent, but $P(A_1 \cap A_2 \cap A_3) = P(A_1)\, P(A_2)\, P(A_3)$ is satisfied.

So, to get around this problem we add some more conditions and get the following definition.

Definition 5 : Three events A_1, A_2 and A_3 corresponding to the same random experiment are said to be **stochastically** or **mutually independent** if

$$P(A_1 \cap A_2) = P(A_1)\,P(A_2)$$
$$P(A_2 \cap A_3) = P(A_2)\,P(A_3) \qquad \ldots (15)$$
$$P(A_3 \cap A_1) = P(A_3)\,P(A_1)$$

and $$P(A_1 \cap A_2 \cap A_3) = P(A_1)\,P(A_2)\,P(A_3).$$

Let's try to understand this through an example.

Example 20 : An unbiased coin is tossed three times. Let A_j denote the event that a head turns up on the j-th toss, j = 1, 2, 3. Let's see if A_1, A_2 and A_3 are independent.

Since the coin is unbiased, we assign the same probability, 1/8, to each of the eight possible outcomes.

Check that

$$P(A_1) = P(A_2) = P(A_3) = 1/2$$
$$P(A_1 \cap A_2) = P(A_2 \cap A_3) = P(A_3 \cap A_1) = 1/4, \text{ and}$$
$$P(A_1 \cap A_2 \cap A_3) = 1/8.$$

Thus, all the four equations in (15) are satisfied and the events A_1, A_2, A_3 are mutually independent.

We have seen that the last condition in (15) alone is not enough, since it does not guarantee the independence of pairs of events.

Similarly, the first three equations of (15) alone are not sufficient to guarantee that all the four conditions required for mutual independence would be satisfied. To see this, consider the following example.

Example 21 : An unbiased die is rolled twice. Let A_1 denote the event "odd face on the first roll", A_2 denote the event "odd face on the second roll" and A_3 denote the event that the total score is odd. With the classical assignment of probability 1/36 to each of the sample points, you can easily check that

$$P(A_1) = P(A_2) = P(A_3) = 18/36 = 1/2, \text{ and that}$$
$$P(A_1 \cap A_2) = P(A_2 \cap A_3) = P(A_3 \cap A_1) = 9/36 = 1/4.$$

Thus, the first three equations in (15) are satisfied. But the last one is not valid. The reason for it is that $P(A_1 \cap A_2 \cap A_3)$ is zero (Do you agree ?), and $P(A_1)$, $P(A_2)$, $P(A_3)$ are all positive.

If the first three equations of (15) are satisfied, we say that A_1, A_2 and A_3 are **pairwise independent.** Example 21 shows that pairwise independence does not guarantee mutual independence.

Now we are sure you can define the concept of independence of n events. Does your definition agree with Definition 6?

Definition 6 : The n events $A_1, A_2, \ldots, A_n$ corresponding to the same random experiment are mutually independent if for all r = 2. . . ., n, $1 \le i_1 < i_2 < \ldots < i_r \le n$, the product rule

$$P(A_{i_1} \cap \ldots \cap A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j}) \qquad \ldots (17)$$

holds.

If n events are independent then any r, $2 \le r \le n$ events out of them should also be independent.

Since r of the n events can be chosen in $\binom{n}{r}$ ways, (17) represents

$$\binom{n}{2} + \binom{n}{3} + \ldots + \binom{n}{n} = 2^n - n - 1$$

conditions.

Try to write Definition 6 for n = 3 and see if it matches Definition 5.

We have already seen that if A_1 and A_2 are independent, then A_1^c and A_2 or A_1 and A_2^c or A_1^c and A_2^c are independent. We now give a similar remark about n independent events.

Remark 4 : If $A_1, A_2, \ldots, A_n$ are n independent events, then we may replace some or all of them by their complements without losing independence. In particular, when $A_1, A_2, \ldots, A_n$ are independent, the product rule (17) holds even with some or all of $A_{i_1}, \ldots, A_{i_r}$ are replaced by their complements.

We shall not prove this assertion, but shall use it in the following examples.

Example 22 : Suppose A_1, A_2, A_3 are three independent events, with $P(A_j) = P_j$ and we want to obtain the probability that at least one of them occurs.

We want to find $P(A_1 \cup A_2 \cup A_3)$. Recall that (Example 8)

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_2 \cup A_3) \\ &\quad - P(A_3 \cap A_1) + P(A_1 \cap A_2 \cap A_3) \\ &= P_1 + P_2 + P_3 - P_1P_2 - P_2P_3 - P_3P_1 + P_1P_2P_3 \\ &= 1 - (1 - P_1)(1 - P_2)(1 - P_3). \end{aligned}$$

We could have arrived at this expression more easily by using Remark 4. This is how we can proceed.

$$\begin{aligned} (A_1 \cup A_2 \cup A_3) &= 1 - P((A_1 \cup A_2 \cup A_3)^c) \\ &= 1 - P(A_1^c \cap A_2^c \cap A_3^c) \\ &= 1 - P(A_1^c)\, P(A_2^c)\, P(A_3^c) \text{ by virtue of Remark 4.} \end{aligned}$$

Example 23 : If A_1, A_2 and A_3 are independent events, then can we say that $A_1 \cup A_2$ and A_3 are independent? Let's see.

We have

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1)\, P(A_2) \end{aligned}$$

$$\begin{aligned} \text{and } P((A_1 \cup A_2) \cap A_3) &= P((A_1 \cap A_3) \cup (A_2 \cap A_3)) \\ &= P(A_1 \cap A_3) + P(A_2 \cap A_3) - P(A_1 \cap A_2 \cap A_3) \\ &= \{P(A_1) + P(A_2) - P(A_1)\, P(A_2)\} P(A_3) \\ &= P(A_1 \cup A_2)\, P(A_3), \end{aligned}$$

implying the independence of $A_1 \cup A_2$ and A_3.

Example 24 : An automatic machine produces bolts. Each bolt has probability 1/10 of being defective. Assuming that a bolt is defective independently of all other bolts, let's find

i) the probability that a good bolt is followed by two defective ones.

ii) the probability of getting one good and two defective bolts, not necessarily in that order.

Let A_j denote the event that the j-th inspected bolt is defective, j = 1, 2, 3. The assumption of independence implies that A_1, A_2 and A_3 are independent.

i) We want $P(A_1^c \cap A_2 \cap A_3)$. By Remark 4, we can write

$$P(A_1^c \cap A_2 \cap A_3) = P(A_1^c)\, P(A_2)\, P(A_3)$$

$$= \frac{9}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} = 0.009.$$

ii) We want to find the probability of

$$(A_1^c \cap A_2 \cap A_3) \cup (A_1 \cap A_2^c \cap A_3) \cup (A_1 \cap A_2 \cap A_3^c).$$

Notice that these events are disjoint and that each has the probability 0.009 (see (i)). Hence, the required probability is

$$P(A_1^c \cap A_2 \cap A_3) + P(A_1 \cap A_2^c \cap A_3) + (A_1 \cap A_2 \cap A_3^c)$$

$$= 3 \times 0.009 = 0.027.$$

Example 25 : The probability that a person A will be alive 20 years hence is 0.7 and the probability that another person B will be alive 20 years hence is 0.5. Assuming independence, let's find the probability that neither of them will be alive after 20 years.

The probability that A dies before twenty years have elapsed is 0.3 and the corresponding probability for B is 0.5. Hence the probability that neither of them will be alive 20 years hence is

$$0.3 \times 0.5 = 0.15,$$

by virtue of independence.

We now give you some exercises based on the concept of independence.

E21) If A_1, A_2 and A_3 are independent events, examine for independence the following pairs of events :

a) A_1 and $A_2 \cap A_3$

b) A_1 and $A_2^c \cup A_3^c$

c) A_1^c and $A_2^c \cap A_3^c$.

E22) Obtain the probabilities of

a) $A_1 \cup (A_2 \cap A_3)$

b) $A_1 \cap (A_2^c \cap A_3^c)$

c) $A_1^c \cap (A_2^c \cap A_3^c)$

under the assumptions of E21, if

$P(A_1) = P(A_2) = P(A_3) = 1/3$.

E23) Suppose that a sample space Ω consists of six permutations of (a, b, c) and three additional points (a, a, a), (b, b, b) and (c, c, c). Each one of the nine points is assigned the probability 1/9. Let A_k denote the event that k-th place is occupied by the letter c, $k = 1, 2, 3$. Are A_1, A_2 and A_3 mutually independent events?

E24) Let A_1, A_2, A_3 and A_4 be four independent events with the same probability 1/3. Obtain the probability that exactly two of them occur.

Hint : You have to first find the probabilities

$P(A_1 \cap A_2 \cap A_3^c \cap A_4^c)$, $P(A_1 \cap A_2^c \cap A_3^c \cap A_4)$

$P(A_1 \cap A_2^c \cap A_3 \cap A_4^c)$, $P(A_1^c \cap A_2^c \cap A_3 \cap A_4)$

$P(A_1^c \cap A_2 \cap A_3^c \cap A_4)$, $P(A_1^c \cap A_2 \cap A_3 \cap A_4^c)]$

E25) Let $\Omega = \{(a, a), (a, b), (b, a), (b, b)\}$. Let A_k be the event that letter 'a' appears at the k-th place, $k = 1, 2$. Examine A_1 and A_2 for independence under the following assignments of probabilities.

		Sample point			
		(a, a)	(a, b)	(b, a)	(b, b)
Assignment	1	1/4	1/4	1/4	1/4
	2	1/18	5/18	1/2	1/6

In E25 you must have found that A_1 and A_2 are independent under Assignment 1 but not under Assignment 2. This shows that independence of events depends on the assignment of probabilities to the sample points and is not their intrinsic property.

The discussion so far has related to a random experiment performed only once. But usually scientists carry out the same experiment more than once and preferably under identical conditions. In the next section, we shall consider the extension of our study to cover such cases which involve repetition of an experiment or which involve performing two or more distinct experiments.

6.6 REPEATED EXPERIMENTS AND TRIALS

We must mention that we have earlier discussed rolls of two dice or three or more tosses of a coin without bringing in the concept of repeated trials. The following discussion is only an elementary introduction to the topic of repeated trials.

To fix ideas, consider the simple experiment of tossing a coin twice. The sample space corresponding to the first toss is $S_1 = \{H, T\}$ say, where H = Head, T = Tail. Similarly the sample space S_2 for the second toss is also $\{H, T\}$. Now observe that the sample space for two tosses is $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$, where (H, H) stands for head on first toss followed by a head on the second toss. The pairs (H, T), etc. are also similarly defined. Note that Ω consists of all **ordered** pairs that can be formed by choosing a point from S_1 followed by a point from S_2. Mathematically we say that Ω is the Cartesian product $S_1 \times S_2$ (read, S_1 cross S_2) of S_1 and S_2.

Now consider an experiment of tossing a coin and then rolling a die. The sample space corresponding to toss of the coin is $S_1 = \{H, T\}$ and that corresponding to the roll of the die is $S_2 = \{1, 2, 3, 4, 5, 6\}$. The sample space of the combined experiment is

$$\Omega = \{(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6),$$
$$(T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\} = S_1 \times S_2.$$

Taking a cue from these two examples we can say that if S_1 and S_2 are the sample spaces for two random experiments ε_1 and ε_2, then the Cartesian product $S_1 \times S_2$ is the sample space of the experiment consisting of both ε_1 and ε_2.

Sometimes we refer to $S_1 \times S_2$ as the **product space** of the two experiments.

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

E26) Find the sample spaces of the following experiments

a) Rolling two dice

b) Drawing two cards from a pack of 52 playing cards, with replacement.

Do you remember the definition of the Cartesian product of $n (n \geq 3)$ sets? We say that the Cartesian product

$$S_1 \times S_2 \times \ldots \times S_n = \{(x_1, \ldots, x_n) \mid x_j \in S_j, j = 1, \ldots n\}.$$

Now, if $S_1, S_2, \ldots, S_n$ represent the sample spaces corresponding to repetitions $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ of the same experiment ε, then the Cartesian product $S_1 \times S_2 \times \ldots \times S_n$ represents the sample space for n repetitions or n trials of the experiment ε.

We now return to the experiment of two tosses of a coin. The sample space is $\Omega = \{(H, H),$

(H, T), (T, H), (T, T)} which is the Cartesian product of {H, T} with itself. Suppose the coin is unbiased so that P{H} = P{T} = 1/2 for both the first and the second toss. Since the coin is unbiased, we may regard the four points in Ω as equally likely and assign probability 1/4 to each one of them. However, another way of looking at this assignment is to assume that the results in the two tosses are independent. More specifically, we may consider specifying P{(H, H)}, say, by the multiplication rule available to us under independence, i.e., we may take

$$P\{(H, H)\} = P\{H\}.\ P\{H\} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

and make similar calculations for other points.

When such a situation holds, we say that the two tosses or the two trials of tossing the coin are independent. This is equivalent to saying that the events Head on first toss and Head on second toss are independent and that we may make similar statements about the other points also. The following example illustrates the method of defining probabilities on the product spaces when we are unable (or unwilling) to assume equally likely outcomes.

Example 26 : Suppose the successive units manufactured by a machine are such that each unit has probability p of being defective (D) and (1 – p) of being good (G). We examine three units manufactured by this machine. The sample space for this experiment is the Cartesian product $S_1 \times S_2 \times S_3$, where $S_1 = S_2 = S_3 = \{D, G\}$, i.e.

$$\Omega = \{(D, D, D), (D, D, G), (D, G, D), (G, D, D),$$
$$(G, G, D), (G, D, G), (D, G, G), (G, G, G)\}.$$

The statement that "the successive units are independent of each other" is interpreted by assigning probabilities to points of Ω by the product rule. In particular,

$$P\{(D, D, D)\} = P\{D\}\ P\{D\}\ P\{D\} = P^3,$$
$$P\{(D, D, G)\} = P\{D\}\ P\{D\}\ P\{G\} = P^2q$$
$$= P\{(D, G, D)\} = P\{(G, D, D)\},$$
$$P\{(G, G, D)\} = P\{G\}\ P\{G\}\ P\{D\} = (1 - p)^2p$$
$$= P\{(G, D, G)\} = P\{(D, G, G)\},$$

and lastly,

$$P\{(G, G, G)\} = P\{G\}\ P\{G\}\ P\{G\} = (1 - p)^3.$$

Notice that the sum of the probabilities of the eight points in Ω is

$$p^3 + 3p^2(1 - p) + 3p(1 - p)^2 + (1 - p)^3$$
$$= \{p + (-p)\}^3 = 1,$$

which is as it should be.

Summarising the discussion so far, consider two random experiments ε_1 and ε_2 with sample spaces S_1 and S_2, respectively. Let $u_1, u_2, \ldots$ be the points of S_1 and let $v_1, v_2 \ldots$, be the points of S_2. Suppose $p_1, p_2, \ldots$, and $q_1, q_2, \ldots$ are the associated probabilities, i.e., $P\{u_i\} = p_i$ and $P\{v_j\} = q_j$, with $p_i, q_j \geq 0$, $\sum_i p_i = 1$, $\sum_j q_j = 1$. We say that ε_1 and ε_2 are independent experiments if the events "first outcome is u_i" and the event "second outcome is v_j", are independent,

i.e., if the assignment of probabilities on the product space $S_1 \times S_2$ is such that

$$P\{(u_j, v_j)\} = P\{u_i\}\ P\{v_j\} = p_i\, q_j.$$

This assignment is a valid assignment because $P\{(u_i, v_j)\} \geq 0$ and

$$\sum_i \sum_j P\{(u_i, v_j)\} = \sum_i \sum_j p_i\, q_j$$
$$= \sum_i p_i \sum_j q_j = 1,$$

where the sums are taken over all values of i and j.

Can we extend these concepts to the case of n {n > 2} random experiments?

Let us denote the n random experiments by $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Let $S_1, S_2, \ldots, S_n$ be the corresponding sample spaces. Let $P\{x_j\}$ denote the probability assigned to the outcome x_j of the random experiment ε_j. We say that $\varepsilon_1, \ldots, \varepsilon_n$ are independent experiments, if the assignment of probabilities on the product space $S_1 \times S_2 \times \ldots \times S_n$ is such that

$$P\{(x_1, x_2, \ldots, x_n)\} = P\{x_1\} P\{x_2\} \ldots P\{x_n\}.$$

The random experiments $\varepsilon_1, \ldots, \varepsilon_n$ are said to be repeated independent trials of an experiment ε if the sample space of $\varepsilon_1, \ldots, \varepsilon_n$ are all identical and so are the assignment of probabilities, it is in this sense that the experiment discussed in Example 26 corresponds to 3 independent repetitions of the experiment of inspecting a unit, where the probability of a unit being defective is P.

Before we conclude our discussion of product spaces and repeated trials, let us revert to the case of two independent experiments ε_1 and ε_2 with sample spaces S_1 and S_2.

Suppose

$$S_1 = \{u_1, u_2, \ldots\}, P\{u_i\} = p_i, i \geq 1$$

$$S_2 = \{v_1, v_2, \ldots\}, P\{v_j\} = q_j, j \geq 1.$$

Let $A_1 = \{u_{i_1}, u_{i_2}, \ldots\}$ and $A_2 = \{v_{j_1}, v_{j_2}, \ldots\}$ be two events in S_1 and S_2. Then $A_1 \times A_2$ is an event in $S_1 \times S_2$ and

$$A_1 \times A_2 = \{(u_{i_r}, v_{j_s}) \mid r, s = 1, 2, \ldots\}.$$

Under the assumption that ε_1 and ε_2 are independent, we can write

$$P\{A_1 \times A_2\} = \sum_r \sum_s P\{(u_{i_r}, v_{j_s})\}$$

$$= \sum \sum p_i q_j$$

$$= \sum_r p_{i_r} \sum_s q_{j_s}$$

$$= P(A_1) \times P(A_2).$$

Thus, the multiplication rule is valid not only for individual sample points of $S_1 \times S_2$ but also for events in the component sample spaces S_1 and S_2 also. Here we have talked about events related to two experiments. But we can easily extend this fact to events related to three or more experiments.

Can you see the parallel between three independent Bernoulli trials and the situation in Example 26?

The independent Bernoulli trials provide the simplest example of repeated independent trials. Here each trial has only two possible outcomes, usually called success (S) and failure (F). We further assume that the probability of success is the same in each trial, and therefore, the probability of failure is also the same for each trial. Usually we denote the probability of success by p and that of failure by $q = 1 - p$.

Suppose, we consider three independent Bernoulli trials. The sample space is the Cartesian product $\{S, F\} \times \{S, F\} \times \{S, F\}$. It, therefore, consists of the eight points

SSS, SSF, SFS, FSS, FFS, FSF, SFF, FFF.

In view of independence, the corresponding probabilities are

$$p^3, p^2q, p^2q, p^2q, pq^2, pq^2, pq^2, q^3.$$

Do they add up to one? Yes.

In general, the sample space corresponding to n independent Bernoulli trials consists of 2^n points. A generic point in this sample space consists of the sequence of n letters, j of which are S and $n - j$ are F, $j = 0, 1, \ldots, n$. Each such point carries the probability $p^j q^{n-j}$, irrespective of the arrangement of j S's and $(n - j)$ F's. Suppose we want to find the

probability of j successes in n independent Bernoulli trials. We first note that there are $\binom{n}{j}$ points with j successes and (n – j) failures (we ask you to prove this in E27). Since each such point carries the probability $p^j q^{n-j}$, the probability of j successes, denoted by b(j, n, p) is

$$b(j, n, p) = \binom{n}{j} p^j q^{n-j}, \; j = 0, 1, \ldots, n.$$

These are called binomial probabilities and we shall return to a discussion of this topic when we discuss the binomial distribution in Unit 8.

E27) Prove that there are $\binom{n}{j}$ points with j successes and (n – j) failures in n Bernoulli trials.

Now we bring this unit to a close. But before that let's briefly recall the important concepts that we studied in it.

6.7 SUMMARY

In this rather lengthy unit, we discussed the following main points :

1) We have introduced you to the axiomatic approach to the definition of probability. In this approach we assign probabilities $P(\omega_j)$ to the points of a discrete sample space

$$\Omega = \{\omega_1, \omega_2, \ldots\}$$

such that

i) $0 \leq P(\omega_j) \leq 1, j = 1, 2, \ldots$

ii) $\sum_j P(\omega_j) = 1.$

2) We have seen how to compute the probability of an event A and have discussed its various properties.

3) We have noted that the classical definition of probability assigns equal probabilities to each of the points of a finite sample space.

4) We have acquainted you with the concept of conditional probability P(A | B) of a given the event B.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \; P(B) > 0.$$

5) We have stated and proved Bayes' theorem :

If $B_1, B_2, \ldots, B_n$ are n events which constitute a partition of Ω, and A is an event of positive probability, then

$$P(B_r \mid A) = \frac{P(B_r)\, P(A \mid B_r)}{\sum_1^n P(B_j)\, P(A \mid B_j)},$$

for any r, $1 \leq r \leq n$.

6) We have defined and discussed the consequences of independence of two or more events.

7) We have seen the method of assignment of probabilities when dealing with independent repetitions of an experiment.

6.8 SOLUTIONS AND ANSWERS

E1) a) and e) are valid.

b) is not valid as the sum of the probabilities is less than one.

c) is not valid since $P\{\omega_8\} < 0$.

d) is not valid since the sum of all the probabilities is greater than one.

E2) a) $0 \le P(\omega_i) \le 1$ for $i = 1, 2, \ldots, 8$ and

$$\sum_{1}^{8} P(\omega_i) = 1 \text{ (using binomial theorem).}$$

Therefore, the assignment is valid.

b) i) The probability of finding exactly one bad item

$$= P\{\omega_2, \omega_3, \omega_4\}$$

$$= 3.\left(\frac{9}{10}\right)^2\left(\frac{1}{10}\right)$$

$$= \frac{243}{1000}.$$

ii) P{GGG, GGB, GBG, BGG, BBG, BGB, BBG}

$$= \left(\frac{9}{10}\right)^3 + 3\left(\frac{9}{10}\right)^2\left(\frac{1}{10}\right) + 3\left(\frac{9}{10}\right)\left(\frac{1}{10}\right)^2$$

$$= \frac{999}{1000}.$$

E3) If A and B are disjoint, $A \cap B = \phi$ and P3 follows from P2.

$A \cap A^c = \phi$ and $A \cup A^c = \Omega$

Hence $P(A \cup A^c) = P(\Omega) = 1$

$P(A) + P(A^c) = 1$ by P3.

P4 follows.

P5 is a simple consequence of P1 and P2.

E4) a) Use P6 to claim $P(A \cup B) = 1$, and write $P(A \cap B) = 1 - P(A^c \cup B^c)$ to obtain $P(A \cap B) = 1$.

b) Use Boole's inequality

c) $P(\phi) = P(\Omega^c) = 1 - P(\Omega) = 0$.

E5) Use $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

P(A)	P(B)	P(A ∪ B)	P(A ∩ B)
0.4	0.8	**0.9**	0.3
0.35	0.5	0.6	0.25

E6) a) violates P4

b) violates P5

c) and d) violate the Axiom.

E7) Let S and W also denote the events that they are absent.

Then $P(S) = 0.05$, $P(W) = 0.10$, $P(S \cap W) = 0.02$. Then a) $P(S^c \cap W^c) = 0.87$, b) $P(S^c \cup W^c) = 0.98$ and c) $P(S \cap W^c) + P(S^c \cap W) = 0.03 + 0.08 = 0.11$.

E8) Use result (5) in Example 4 to obtain the required probability which is 0.80.

E9) $\frac{4 \times 4}{52 \times 52} \cong 0.0059$.

E10) a) $\binom{8}{3} / \binom{10}{3} \cong 0.467$.

b) $\binom{5}{2}\binom{3}{1} / \binom{10}{3} \cong 0.25$.

E11) Let p_1, p_2, p_3 and p_4 denote the probabilities that the number of attempts is 1, 2, 3, and 4, respectively. Then,

$$p_1 = 1/4,\ p_2 = \frac{3\times 1}{4\times 3} = \frac{1}{4},\ p_3 = \frac{3\times 2\times 1}{4\times 3\times 2} = \frac{1}{4}$$

and $p_4 = \frac{3\times 2\times 1\times 1}{4\times 3\times 2\times 1} = 1/4.$

E12) a) $80/200 = 0.4$

b) $60/200 = 0.3$

c) $120/200 = 0.6.$

E13) $\frac{(4+3+2+1)}{36} \cong 0.278.$

E14) You should first interpret A^c and B^c and then explain. For example, b) is the probability that a randomly selected person does not suffer from high blood pressure given that he/she is a smoker.

E15) Required conditional probability $= \frac{1/36}{6/36} = 1/6.$

E16) a) $1 - 0.4 = 0.6$

b) $0.6 \times 0.6 = 0.36$

c) $0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \cong 0.168.$

E17) Required probability $= (1/2)\ (4/10) + (1/2)\ (6/10) = 0.5.$

E18) By Bayes' theorem, the required probability is

$$\frac{0.02 \times 0.80}{0.02 \times 0.80 + 0.98 \times 0.05} \cong 0.246.$$

E19) Let A_1, A_2, A_3 and A_4 denote the four causes of explosion and E denote the event of explosion. We need to compute $P(A_1 \mid E)$, $P(A_2 \mid E)$, $P(A_3 \mid E)$ and $P(A_4 \mid E)$. We have $P(A_1) = 0.20$, $P(A_2) = 0.30$, $P(A_3) = 0.40$, $P(A_4) = 0.10$ and $P(E \mid A_1) = 0.25$, $P(E \mid A_2) = 0.20$, $P(E \mid A_3) = 0.50$, $P(E \mid A_4) = 0.75$.

Using Bayes' theorem, we get

$P(A_1 \mid E) = 0.181$, $P(A_2 \mid E) = 0.218$

$P(A_3 \mid E) = 0.327$, $P(A_4 \mid E) = 0.273.$

Thus, the most likely cause of explosion is the carelessness of workers.

E20) a) $P(A_1) = P(A_2) = \frac{18}{36} = \frac{1}{2}$

and $P(A_1 \cap A_2) = \frac{9}{36} = \frac{1}{4}$

$\therefore\ P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$, and hence, A_1 and A_2 are independent.

b) $P(B_1) = \frac{1}{6}$, $P(B_2) = \frac{1}{12}$,

$P(B_1 \cap B_2) = \frac{1}{18}.$

B_1 and B_2 are not independent.

E21) The pairs in all the three cases are independent. You need to verify that the product rule holds. This is obvious in a). In establishing b), use P2 and P4. The result in c) follows on using $P(A_1^c \cap A_2^c \cap A_3^c = 1 - P(A_1 \cup A_2 \cup A_3)$, result (5) in Example 4, followed by algebraic simplication.

E22) a) 11/27

b) 8/27

c) 8/27

E23) Verify that $P(A_1) = P(A_2) = P(A_3) = 1/3$.

$P(A_1 \cap A_2) = 1/9 = P(A_2 \cap A_3) = P(A_3 \cap A_1)$ and that $P(A_1 \cap A_2 \cap A_3) = 1/9$. Thus A_1, A_2 and A_3 are not mutually independent. In fact, this is one more example of pairwise independence not implying independence.

E24) Each of the probabilities given in **Hint** is 4/81. The required probability is the sum of all these six probabilities, i.e., $\frac{24}{81}$.

E25) $A_1 = \{(a, a), (a, b)\}$

$A_2 = \{(a, a), (b, a)\}$

$\therefore A_1 \cap A_2 = \{(a, a)\}$

$P(A_1 \cap A_2 = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_1) \cdot P(A_2)$ in Assignment 1.

$P(A_1 \cap A_2) \neq P(A_1) \cdot P(A_2)$ in Assignment 2.

E26) a) Let S be the sample space of the experiment of rolling a die.

$\therefore S = \{1, 2, 3, 4, 5, 6\}$.

The required sample space = $S \times S$.

b) The required sample space is $S \times S$, where S = {52 playing cards}.

E27) We have mentioned that a point in the sample space of n Bernoulli trials is a sequence of n letters, j of which are S and the remaining n – j are F, j = 1, 2, . . . , n.

Now consider the situation where we have n vacant slots and j copies of the letter S. In how many ways can we put these j letters in the n slots ? You know that this can be done in $\binom{n}{j}$ ways. Once the Ss are in place, the remaining vacant slots can be filled with Fs.

So, in each of these $\binom{n}{j}$ ways, we have j, Ss and (n – j), Fs occupying the n slots.

UNIT 7 DISCRETE RANDOM VARIABLE AND ITS PROBABILITY DISTRIBUTION

Structure

7.1 INTRODUCTION

We have seen a number of examples of sample spaces of random experiments in the previous two units. You must have noticed that in most practical applications a numerical value is associated with each outcome of a random experiment. Mathematically speaking, we have a real-valued function defined on the sample space. Such a function is called a random variable. This unit is devoted to the study of a random variable, defined on a discrete sample space. We introduce the concept of such a random variable and its probability distribution in Sec. 7.2. In Sec. 7.3 we describe the joint probability distribution of two or more random variables which leads to a discussion of marginal distributions and independence of random variables. The mathematical expectation (mean), variance of a random variable, covariance and correlation of two random variables are discussed in Sections 7.5 and 7.6, respectively. Then we generalise these concepts to introduce moments and moment generating functions. You have already come across these terms in the context of a frequency distribution in Block 1. Here we are going to discuss them in the context of a discrete probability distribution. We conclude this unit with an introduction to the problem of obtaining the distribution of the sum of two random variables.

Objectives

A study of this unit would enable you to :

- define a random variable and specify its probability distribution,
- specify the joint distribution of two or more random variables,
- obtain their marginal distributions and examine them for their independence,
- define and calculate the means, variances, covariances and correlation coefficients of random variables,
- define moments and obtain moment generating functions,
- obtain the probability distribution of the sum of two random variables.

7.2 RANDOM VARIABLE

In the first two units of this block we have introduced the concepts of a random experiment, associated sample space and probability of an event. With the help of these we study the uncertainties associated with such experiments. We usually find that a numerical measurement or quantity is associated with a random experiment. Consider the following examples :

1) A person invests Rs. 1 in purchasing a lottery ticket. He either wins the first prize of Rs. 100 or loses his rupee. His net gain is either –1 or 99. This net gain cannot be predicted in advance.

2) The authorities of IGNOU cannot predict in advance the number of students who would join and complete this course. This number could be 0, 1, 2, ...

3) The number of calls that a telephone exchange would receive in a specified time interval can be 0, 1, 2, ...

4) The total number of defects in a motor cycle coming off a production line can be any number like 0, 1, 2, ...

5) The maximum temperature of Delhi on June 05, can be anywhere between 40° and 50° C.

All these examples have one common feature. They describe a numerical characteristic associated with a random experiment. This characteristic depends on the outcome of the experiment and therefore its value cannot be predicted in advance.

The numerical characteristic associated with a random experiment is a variable quantity which behaves randomly and so we may call it a "random variable". This is of course, not a technical definition of the term "random variable".

In order to make our ideas precise, we consider an example. Suppose we are interested in the number X of heads obtained in three tosses of a coin.

The sample space Ω consists of the eight points

$$\omega_1 = HHH,\ \omega_2 = HHT,\ \omega_3 = HTH,\ \omega_4 = THH,\ \omega_5 = TTH,\ \omega_6 = THT,$$

$$\omega_7 = HTT,\ \omega_8 = TTT.$$

We could be also interested in the number X of girls in families with three children.

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

$$X(\omega_1) = 3,\ X(\omega_2) = X(\omega_3) = X(\omega_4) = 2,$$
$$X(\omega_5) = X(\omega_6) = X(\omega_7) = 1,\ X(\omega_8) = 0$$

Do you agree that, the number X of heads in three tosses of a coin is a **function** defined on the sample space Ω? It assumes the values 0, 1, 2 and 3, as you have been above. Observe, now that

X = k means that there are k heads in the outcome.

$$X = \begin{cases} 0 \text{ iff the outcome is } \omega_8 \\ 1 \text{ iff the outcome is } \omega_5, \omega_6 \text{ or } \omega_7 \\ 2 \text{ iff the outcome is } \omega_2, \omega_3, \text{ or } \omega_4 \\ 3 \text{ iff the outcome is } \omega_1. \end{cases}$$

We can, therefore, make the following identification of events:

[X = k], k = 0, 1, 2, 3, is a subset of Ω, and hence is an event.

$$[X = 0] = \{\omega_8\},\ [X = 1] = \{\omega_5, \omega_6, \omega_7\}$$

$$[X = 2] = \{\omega_2, \omega_3, \omega_4\},\ [X = 3] = \{\omega_1\}.$$

Suppose now that

$$P\{\omega_1\} = P\{\omega_2\} = \ldots = P\{\omega_8\} = 1/8.$$

Then because of the above identification of events [X = j], j = 0, 1, 2, 3, we can write

$$P[X = 0] = P\{\omega_8\} = 1/8,\ P[X = 1] = P\{\omega_5, \omega_6, \omega_7\} = P\{\omega_5\} +$$

$$P\{\omega_6\} + P\{\omega_7\} = 3/8,$$

$$P[X = 2] = 3/8$$

and $P[X = 3] = 1/8$.

where we read $P[X = j]$ as "probability that X equals j." Have you noticed that $[X = j]$, $j = 0, 1, 2, 3$ are mutually disjoint sets, and that

$$\bigcup_{j=0}^{3} [X = j] = \Omega ?$$

Also note that

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

which is as it should be (see Axiom in Unit 6).

Now let us sum up and list the essential properties of the number X of heads obtained in three tosses of a coin.

i) X is a function defined on the sample space Ω.

ii) It assumes a finite number of real values.

iii) We can assign a probability to the event that X assumes a particular value.

iv) The sum of the probabilities that X assumes the different values is one.

In this unit (and in this block) we shall restrict our attention to discrete sample spaces. So, on the basis of the above discussion we give the following definition.

Definition 1 : A random variable is a real-valued function on a discrete sample space Ω.

In what follows we shall denote random variables by capital letters, X, Y, W, U, V,..., with or without suffixes. The value of a random variable X at a point ω in the sample space Ω, will be denoted by $X(\omega)$. We shall also write r.v. for random variable.

Recall that a discrete sample space has either a finite number of points or its points can be arranged in a sequence. Since an r.v. is a function on the sample space, it can take either a finite number of values or its values can be arranged in a sequence. Suppose, therefore, that an r.v. X takes the values $x_1, x_2, \ldots$. Denote the probability $P[X = x_j]$ that X takes the value x_j by $f(x_j)$, $j = 1, 2, \ldots$. Then we have the following definition.

Definition 2 : The function $f(x_j) = P[X = x_j]$, $j = 1, 2, \ldots$, defined for the values $x_1, x_2, \ldots$ assumed by X is called the **probability mass function** of X.

Sometimes it is also called the probability distribution of X.

Do you agree that $f(x_j) \geq 0$? What about the sum

$$f(x_1) + f(x_2) + \ldots ?$$

Now X is a function from Ω to **R**. Therefore, the sets $[X = x_j] = \{\omega \in \Omega \mid X(\omega) = x_j\}$, $j = 1, 2, \ldots$, are all mutually disjoint. Because, if

$$\omega \in [X = x_j] \cap [X = x_k] \text{ for some } j \neq k, \text{ then}$$

$$X(\omega) = x_j \text{ and } X(\omega) = x_k, \text{ where } x_j \neq x_k.$$

this is impossible since X is a function.

$$\text{Further, } \bigcup_j [X = x_j] = \Omega.$$

$$\text{Hence, } \sum_j f(x_j) = \sum_j P[X = x_j] = P[\bigcup_j [X = x_j]]$$

$$= P(\Omega) = 1.$$

We now give some examples concerning probability mass functions.

Example 1 : We have seen that the probability mass function of the r.v. X denoting the number of heads obtained in three tosses of a coin is,

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

Example 2 : An unbiased die is rolled twice. Let X denote the total score so obtained. The sample space of this experiment is the set $\Omega = \{(x, y) \mid x, y = 1, \ldots 6\}$ of all ordered pairs (x, y), x being the score obtained on the first throw and y that on the second throw. Each of the 36 points in Ω carries the probability 1/36. Now what values does X take ? X take the values 2, 3, . . . , 12. In the following table we identify the subjects corresponding to the events [X = j], j = 2, 3, . . . , 12, as well as the corresponding probabilites, f(2), . . . f(12).

Table 1 : Probability Mass Function of X

j	Event [X=j]	Subset of Ω	f(j) = P [X = j]
2	[X = 2]	{(1, 1)}	1/36
3	[X = 3]	{(1,2), (2, 1)}	2/36
4	[X = 4]	{(1, 3) (2, 2), (3, 1)}	3/36
5	[X = 5]	{(1, 4) (2, 3), (3, 2), (4, 1)}	4/36
6	[X = 6]	{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)}	5/36
7	[X = 7]	{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)}	6/36
8	[X = 8]	{ (2, 6), (3, 5), (4, 4), (5, 3), (6, 2) }	5/36
9	[X = 9]	{(3, 6), (4, 5), (5, 4), (6, 3)}	4/36
10	[X = 10]	{(4, 6), (5, 5), (6, 4)}	3/36
11	[X = 11]	{(5, 6), (6, 5)}	2/36
12	[X = 12]	{(6, 6)}	1/36

You can see that f(j) > 0 for all j.

You can also check that f(2) + f(3) + . . . f(12) = 1.

Example 3 : An r.v. X takes the values 1, 2, . . . , k, with probabilities

$$P[X = j] = f(j) = cj, \; j = 1 \ldots, k.$$

Let us find the constant c such that f(j) is a probability mass function.

If f is a probability mass function, we must have

$$f(1) + f(2) + \ldots + f(k) = 1.$$

i.e. $c\{1 + 2 + \ldots + k\} = 1,$

i.e., $c\,\frac{k(k+1)}{2} = 1$, implying that $c = \frac{2}{k(k+1)}$. Clearly, c > 0,

which implies that $f(j) > 0 \; \forall_j$. Thus, the probability mass function of X is

$$f(j) = \frac{2j}{k(k+1)}, \; j = 1, \ldots, k.$$

Now you can extend the arguments used in Example 3 to solve this exercise.

E1) An r.v. X takes the values 0, 1, 2, . . . with probabilities

$$f(j) = cp^j, \; j = 0, 1, 2, \ldots,$$

where $0 < p < 1$. Determine c such that f(j) is a probability mass function.

In E1 you must have seen that f(j)s are terms of a convergent geometric series. Therefore, we say that the r.v. X with the probability mass function in E1 has a **geometric distribution.**

Let us return to the discussion of the three tosses of an unbiased coin. The r.v. X, denoting the number of heads so obtained, has the probability mass function

$$f(0) = \frac{1}{8}, \; f(1) = \frac{3}{8}, \; f(2) = \frac{3}{8}, \; f(3) = \frac{1}{8}.$$

Suppose we want to know the probability $P[X \leq 2]$. Since $X \leq 2$ iff, X = 0 or 1 or 2, and since the events [X = 0], [X = 1] and [X = 2] are disjoint, we can write,

$$P[X \leq 2] = P[X = 0] + P[X = 1] + P[X = 2]$$

$$= f(0) + f(1) + f(2)$$

$$= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}.$$

Similarly, we can obtain the probability that the sum of the scores obtained by rolling a die twice is greater that 8. In fact, for r.v. X of Example 2,

$$P[X > 8] = P[X = 9] + P[X = 10] + P[X = 11] + P[X = 12]$$

$$= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36}$$

$$= \frac{5}{18}.$$

More generally, let H be any subset of the set of possible values of an r.v. X. Then

$$P[X \in H] = \sum_{x_j \in H} f(x_j).$$

using the properly P7 or P8 of Unit 6. Here the sum is taken over all points x_j in the subset H.

Suppose we have a random variable X assuming values $x_1, x_2, \ldots$ with probabilities $f(x_1)$, $f(x_2), \ldots$, respectively. You may also visualise this as an illustration of a frequency distribution. The values $x_1, x_2, \ldots$ assumed by the random variable correspond to the values of the variable or to mid-values of the class-intervals, and the probabilities $f(x_1), f(x_2), \ldots$ play the role of relative frequencies. We will find this interpretation useful when studying expectation and variance of a random variable.

In what follows, we shall study the properties of a random variable only in terms of its probability mass function. That is, we may not always refer to the underlying sample space or to the specification of the function on the sample space which yields random variable with specified probability mass function. However, we can always visualise a random experiment which leads to a random variable with specified probability mass function. To see this, imagine a box containing cards bearing the numbers $x_1, x_2 \ldots$, and let $f(x_j)$ be the proportion of cards bearing the number x_j, $j = 1, 2, \ldots$. If we choose one of the cards **at random** from this box, then it will bear the number x_j with probability $f(x_j)$, $j = 1, 2, \ldots$. Thus, we have a random experiment which yields a random variable with a specified probability distribution. Did you notice that we said that we can **visualise** a random experiment and not that we can **construct** an experiment? This is because we will not be able to construct the box or any other mechanical device if some or all of probabilities $f(x_1)$, $f(x_2), \ldots$ are irrational numbers or if the discrete random variable takes infinitely many values.

Thus, although for technical reason it is necessary to consider the sample space on which our r.v. is defined, all its properties can be studied with the help of only the probability mass function. In what follows, **we shall use the short form p.m.f. for probability mass function**.

But before we go any further, it is time to do some exercises.

E2) Let X_1 be the score obtained on the first throw and X_2 be the score obtained on the second throw of an unbiased die. Define $W = X_1 - X_2$. Obtain the p.m.f. of W.

(**Hint** : Follow the method of Example 2.)

E3) Three cards are drawn without replacement from a deck of 52 playing cards. Find the p.m.f. of the number Y of spades in the three cards.

E4) A person has 4 keys with which to open a lock. We selects one of the keys at random from the 4 keys on the first attempt. Subsequently, he discards the keys already used and selects one key at random from the remaining keys. He may require 1, 2, 3 or 4 attempts to open the lock. Obtain the probability distribution of the number of attempts.

If you have done these exercises, you would have got a fairly good grasp of p.m.f. Next we study the joint distribution of random variables.

7.3 TWO OR MORE RANDOM VARIABLES

There are many situations where we have to study two or more r.vs. in connection with a random experiment. The following are some examples of such situations.

Recall that you have already come across joint frequency distributions in Unit 4.

i) A store sells two brands, A and B, of tooth-paste. The sales X and Y of brands A and B, respectively, in one week are of interest. Here X and Y are r.vs., both taking values 0, 1, 2 . . .

ii) Let X denote the number of boys born in a hospital in one week and Y that of girls born in the same hospital in the same week. Then X and Y are r.vs., both taking the values 0, 1, 2 . . .

iii) A group of 50 people is vaccinated against a disease and another group of 40 people is not vaccinated. Let X and Y denote the number of people affected by the disease from the two groups. Then X and Y are r.vs. taking values 0, 1 . . . 50, and 0, 1 . . . , 40, respectively.

iv) Suppose we classify the persons according to the day of the week they were born. If $X_1, X_2, \ldots, X_7$ denote the number of students with birthdays on Monday, Tuesday, . . . , Sunday from a class of 100 students, then $X_1, \ldots, X_7$ are r.vs. taking values 0, 1, . . . , 100 subject to the restriction $X_1 + X_2 + \ldots + X_7 = 100$.

We begin this section by describing methods of studying the **joint distribution** of two or more random variables.

7.3.1 Joint Distribution of Random Variables

Let us consider the following artificial example.

Example 4 : A committee of two persons is formed by selecting them at random and without replacement from a group of 10 persons, of whom 2 are mathematicians, 4 are statisticians and 4 are engineers. Let X and Y denote the number of mathematicians and statisticians, respectively, on the committee. The possible values of X are 0, 1, 2, which are also the possible values of Y. Thus, all the ordered pairs (x, y) of the values of X and Y are (0, 0), (9, 1), (0, 2), (1, 0), (1, 1), (2, 0), (1, 2), (2, 1) and (2, 2).

The total number of ways of selecting two persons from a group of 10 persons is $\binom{10}{2} = 45$. Since the persons are selected at random, each of these 45 ways has the same probability, $\frac{1}{45}$. Consider the event [X = 1, Y = 1] that the committee has one mathematician and one statistician. One mathematician can be selected from two in $\binom{2}{1} = 2$ ways and one statistician can be selected from 4 statisticians in $\binom{4}{1} = 4$ ways. Hence the total number of committees with 1 mathematician and 1 statistician is $2 \times 4 = 8$. Thus $P[X = 1, Y = 1] = \frac{8}{45}$.

To obtain the probability of the event [X = 0, Y = 1], observe that if X = 0, Y = 1, this means that 1 statistician is on the committee and no mathematician is on it. Then the other person on the committee has to be one of the 4 engineers. This engineer can be selected in $\binom{4}{1} = 4$ ways. Hence

$$P[X = 0, Y = 1] = \frac{\binom{4}{1}\binom{4}{1}}{45} = \frac{16}{45}$$

Similarly, we can obtain

$$P[X = 0, Y = 0] = \frac{\binom{4}{2}}{45} = \frac{6}{45}$$

$$P[X = 0, Y = 2] = \frac{\binom{4}{2}}{45} = \frac{6}{45}$$

$$P[X = 1, Y = 0] = \frac{\binom{2}{1}\binom{4}{1}}{45} = \frac{8}{45}$$

$$P[X = 2, Y = 0] = \frac{\binom{2}{2}}{45} = \frac{1}{45}$$

Since the committee has only two members, it is obvious that there are no sample points corresponding to the events [X = 1, Y = 2], [X = 2, Y = 1] and [X = 2, Y = 2]. Hence, their probabilities are all equal to zero

We now summarise these calculations in the following table.

Table 2 : P[X = x, Y = y] for x,y = 0, 1, 2.

x \ y	0	1	2
0	6/45	16/45	6/45
1	8/45	8/45	0
2	1/45	0	0

Note that if we denote by f(x, y) the probability P[X = x, Y = y], the function f(x, y) is defined for all pairs (x, y) of values x and y of X and Y, respectively. Moreover,

$$f(x, y) \geq 0$$

and

$$\sum_{x=0}^{2} \sum_{y=0}^{2} f(x, y) = 1.$$

We say that the function f(x, y) is the joint probability mass function of the r.vs. X, Y. More generally, we have the following definition.

Definition 3 : Let X and Y be two r.vs. associated with the same random experiment. Let $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote those of Y. The function $f(x_j, y_k)$ defined for all ordered pairs (x_j, y_k), j, k = 1, 2 . . . by the relation

$$f(x_j, y_k) = P[X = x_j, Y = y_k]$$

is called the **joint probability mass function** of X and Y.

Note that by definition,

$$f(x_j, y_k) \geq 0$$

and

$$\sum_{j} \sum_{k} f(x_j, y_k) = 1.$$

Moreover, we should clarify that $[X = x_j, Y = y_k]$ really stands for the event $[X = x_j] \cap [Y = y_k]$ and that $[X = x_j, Y=y_k]$ is a simplified and accepted way of expressing the intersection of the two events $[X = x_j]$ and $[Y = y_k]$. Notice also that in Example 4, we had used x and y as the arguments of the p.m.f. and in the definition given above we are using x_j and y_k as the arguments. We shall use both notations and trust that it will not cause any confusion.

Now here is an example.

Example 5 : Suppose X and Y are two r.vs. with p.m.f.

f(x, y) = c (x + y), x = 1, 2, 3, 4, and y = 1, 2. What do you think is the value of c ?

c should be such that $c > 0$ and

$$\sum_{x=1}^{4} \sum_{y=1}^{2} f(x, y) = 1 .$$

The left side of the above equation is

$$c \sum_{x=1}^{4} \sum_{y=1}^{2} (x + y) = c \sum_{x=1}^{4} \{(x + 1) + (x + 2)\}$$

$$= c \sum_{x=1}^{4} (2x + 3) = 32c.$$

Hence, $c = 1/32$ and the joint p.m.f. of X, Y is

$$f(x, y) = \frac{(x + y)}{32}, \quad x = 1, 2, 3, 4; \; y = 1, 2.$$

Let us also obtain P[X = 2], P[Y = 1] and P[Y = 2].

Since Y takes the two values 1 and 2, we can write

$$[X = 2] = [X = 2, Y = 1] \cup [X = 2, Y = 2].$$

Moreover, [Y = 1] and [Y = 2] are disjoint events and therefore the events [X = 2, Y = 1] and [X = 2, Y = 2] are also disjoint. Hence,

$$P[X = 2] = P[X = 2, Y = 1] + P[X = 2, Y = 2]$$

$$= \frac{3}{32} + \frac{4}{32} = \frac{7}{32}.$$

Similarly,

$$P[Y = 1] = P[X = 1, Y = 1] + P[X = 2, Y = 1]$$

$$+ P[X = 3, Y = 1] + P[X = 4, Y = 1]$$

$$= \frac{2}{32} + \frac{3}{32} + \frac{4}{32} + \frac{5}{32} = \frac{14}{32}.$$

Now since Y takes only two values 1 and 2.

$$P[Y = 2] = 1 - P[Y = 1] = 1 - \frac{14}{32} = \frac{18}{32}.$$

Note that $P[Y = 1] = \frac{14}{32}$ and $P[Y = 2] = \frac{18}{32}$ specify the p.m.f. of Y when X and Y have the given joint p.m.f. It is called the marginal probability mass function of Y. We will discuss this concept in more detail in the next section.

Example 6 : Let us obtain the conditional probability $P[X = 4 \mid Y = 2]$, that is, the probability that X = 4 given Y = 2 for Example 5.

By definition of the conditional probability,

$$P[X = 4 \mid Y = 2] = \frac{P[X = 4, Y = 2]}{P[Y = 2]}$$

$$= \frac{6/32}{18/32} = \frac{1}{3}.$$

Examples 5 and 6 illustrate that we can obtain probabilities of events associated with r.vs. X and Y by using the joint p.m.f. Hence, as in the case of a single r.v., the joint p.m.f. of X and Y is said to specify the joint probability distribution of X, Y. It is therefore enough to specify the joint p.m.f. of X and Y to answer any question about them.

The concept of joint distribution of two r.vs. is easily extended to that of three r.vs. X, Y and Z. We now need to specify the p.m.f.

$$f(x_j, y_k, z_i) = P[X = x_j, Y = y_k, Z = z_i]$$

for all ordered triples (x_j, y_k, z_i), of values x_j, y_k and z_i of X, Y and Z.

We can now further extend these concepts to more than three r.vs. But we omit the details since, in this course, we shall be mostly dealing with joint distribution of a pair of r.vs. See if you can solve these exercises now

E5) The joint p.m.f. f(x, y) of two r.vs. X and Y is given in the following table.

x \ y	0	1	2	3
0	1/27	3/27	3/27	1/27
1	3/27	6/27	3/27	0
2	3/27	3/27	0	0
3	1/27	0	0	0

a) Obtain (i) $P[X = 2]$, (ii) $P[Y = 0]$ (iii) $P[X = 1, Y \leq 2]$ (iv) $P[X \leq 2, Y = 0]$ (v) $P[X = 2 \mid Y = 0]$.

b) Are the events $[X = 2]$ and $[Y = 0]$ independent ?

c) Calculate $P[X + Y = 4]$.

E6) The joint p.m.f. of two r.vs., X_1 and X_2 is given by

$$f(x_1, x_2) = \frac{10!\,(1/2)^{x_1}\,(1/8)^{x_2}\,(3/8)^{10-x_1-x_2}}{x_1!\,x_2!\,(10-x_1-x_2)!},$$

where $x_1, x_2 = 0, 1, \ldots, 10$, subject to the restriction that $x_1 + x_2 \leq 10$.

Find the following probabilities

a) $P[X_1 = 3]$

b) $P[X_2 \geq 4]$

c) $P[X_1 = 3 \mid X_2 \geq 4]$

d) $P[X_1 = 3, X_2 > 5]$.

Let's turn our attention to marginal distributions now.

7.3.2 Marginal Distributions and Independence

In unit 4 we have discussed the notion of marginal frequency distributions, where we fix one of the variables and study the frequency distribution of the other. We now study the p.m.f. of the marginal distribution. Later we shall use this to define independent random variables.

Let X and Y be r.vs. with values $x_1, x_2, \ldots$ and $y_1, y_2 \ldots$, respectively and joint p.m.f. $f(x_j, y_k) = P[X = x_j, Y = y_k]$.

We define new functions g and h as follows :

$$g(x_j) = \sum_k f(x_j, y_k),\ j = 1, 2, \ldots . \qquad \ldots (1)$$

and

$$h(y_k) = \sum_j f(x_j, y_k),\ k = 1, 2, \ldots . \qquad \ldots (2)$$

In (1), we keep the value x_j of X fixed and sum $f(x_j, y_k)$ over all values y_k of Y. On the other hand, in (2), y_k is kept fixed and $f(x_j, y_k)$ is summed over all values of X. We wish to interpret the function $g(x_j)$ defined for all values, x_j of X and the function $h(y_k)$ defined for all values y_k of Y. Notice that both g and h being sums of non-negative numbers, are themselves non-negative. Further,

$$\sum_j g(x_j) = \sum_j \sum_k f(x_j, y_k) = 1.$$

Thus, $g(x_j)$ has all the properties of a p.m.f. Similarly, you can verify that $h(y_k)$ also has all the properties of a p.m.f. We call these the p.m.f. of the marginal distributions of X and Y, as you can see from the following definition.

Definition 4 : The function $g(x_j)$ defined for all values x_j of the r.v. X by the relation

$$g(x_j) = \sum_k f(x_j, y_k)$$

is called the **p.m.f. of the marginal distribution of X.** Similarly, $h(y_k)$ defined for all the values y_k of the r.v. Y by the relation

$$h(y_k) = \sum_j f(x_j, y_k)$$

is called the **p.m.f. of the marginal distribution of Y.**

Let's try to understand this concept by taking an example.

Example 7 : Let X, Y be two r.vs. with joint p.m.f. f(x, y) defined by the following table.

Table 3 : Joint p.m.f. f(x, y)

x \ y	0	1	2	3	g(x)
0	0	1/6	1/12	1/12	1/3
1	1/24	1/24	1/8	0	5/24
2	5/24	4/24	1/24	1/24	11/24
h(y)	6/24	9/24	6/24	3/24	1

The marginal p.m.f. g(x) of X is obtained by summing all the elements in each of the rows. Similarly, the marginal p.m.f. of Y is obtained by summing all the elements in each of the columns. This procedure is a straightforward consequence of the definition of g(x) and of h(y) when the joint p.m.f. is defined by the above tabular form. In fact, we have

$$g(0) = P[X = 0] = 1/3$$
$$g(1) = P[X = 1] = 5/24$$
$$g(2) = P[X = 2] = 11/24$$

Similarly,

$$h(0) = P[Y = 0] = 6/24$$
$$h(1) = P[Y = 1] = 9/24$$
$$h(2) = P[Y = 2] = 6/24$$
$$h(3) = P[Y = 3] = 3/24$$

In this example, we have $g(x) = P[X = x]$ and $h(y) = P[Y = y]$ for all x and y.

Is it a coincidence ? No.

Notice that in the general situation,

$$g(x_j) = \sum_k f(x_j, y_k) = \sum_k P[X = x_j, Y = y_k]$$

and recall that the events $[X = x_j\ Y = y_k]$ for fixed x_j and different y_k values are disjoint. Hence, by property P7 and P8 of Unit 6,

$$g(x_j) = P\left[\bigcup_k [X = x_j, Y = y_k]\right]$$
$$= P[X = x_j]\ j = 1, 2, \ldots$$

Similarly,

$$h(y_k) = P[Y = y_k], k = 1, 2, \ldots$$

Here is another example.

Example 8 : Let the joint p.m.f. of X and Y be given by $f(x, y) = \frac{x + y}{30}$ for $x = 0, 1, 2, 3$, and $y = 0, 1, 2$.

Then

$$g(x) = \sum_{y} f(x, y) = f(x, 0) + f(x, 1) + f(x, 2)$$

$$= \frac{x}{30} + \frac{x + 1}{30} + \frac{x + 2}{30}$$

$$= \frac{x + 1}{30}, \; x = 0, 1, 2, 3.$$

Similarly,

$$h(y) = \sum_{x} f(x, y)$$

$$= f(0, y) + f(1, y) + f(2, y) + f(3, y)$$

$$= \frac{y}{30} + \frac{1 + y}{30} + \frac{2 + y}{30} + \frac{3 + y}{30}$$

$$= \frac{2y + 3}{15}, \; y = 0, 1, 2.$$

The discussion, so far, tells us that we can determine the marginal p.m.fs. from a knowledge of the joint p.m.f. of the two r.vs. But is it possible to determine the joint p.m.f. from a knowledge of the marginal p.m.fs. ? To answer this, we consider the following two distinct joint p.m.fs. f_1 and f_2 and the corresponding marginal p.m.fs. The first p.m.f. is given by

$$f_1(0, 0) = 1/4, \quad f_1(0, 1) = 1/4$$

$$f_1(1, 0) = 1/8, \quad f_1(1, 1) = 3/8.$$

The corresponding marginal p.m.fs. are

$$g_1(0) = 1/2, \quad g_1(1) = 1/2,$$

$$h_1(0) = 3/8, \quad h_1(1) = 5/8.$$

Now we define the second p.m.f. as

$$f_2(0, 0) = 3/16, \quad f_2(0, 1) = 5/16,$$

$$f_2(1, 0) = 3/16, \quad f_2(1, 1) = 5/16.$$

For this joint p.m.f., the marginal p.m.fs. are

$$g_2(0) = 1/2, \quad g_2(1) = 1/2,$$

$$h_2(0) = 3/8, \quad h_2(1) = 5/8.$$

So, what do we find ? Although the joint p.m.fs. f_1 and f_2 are different, they lead us to the same marginal p.m.fs., $g_1 = g_2$, $h_1 = h_2$. In other words, the marginal distributions of X and Y do not determine their joint distribution uniquely. However, there is one particular situation where this is possible. We now discuss this situation in detail.

Let X and Y be two r.vs. with joint p.m.f. f(x, y) specified in the following table :

Table 4

x \ y	0	1	2	3	g(x)
0	1/12	1/24	1/24	1/6	1/3
1	1/24	1/48	1/48	1/12	1/6
2	1/8	1/16	1/16	1/4	1/2
h(y)	1/4	1/8	1/8	1/2	1

Consider $f(1, 3) = P[X = 1, Y = 3]$ = 1/12

Y = 3] of the events [X = 1] and [Y = 3] is 1/12. But since g(1) = P[X = 1] = 1/6 and h(3) = P[Y = 3] = 1/2, we have the relation

$$P[X = 1, Y = 3] = f(1, 3) = g(1)\, h(3)$$
$$= P[X = 1]\, P[Y = 3].$$

This means that the events [X = 1] and [Y = 3] are independent. In fact, notice that Table 4 is so constructed that

$$P[X = x, Y = y] = f(x, y) = g(x)\, h(y)$$
$$= P[x = x]\, P[Y = y]$$

for all x = 0, 1, 2, and y = 0, 1, 2, 3. In other words, for all possible values x of X and y of Y, the events [X = x] and [Y = y] are independent. In such a situation, we are justified in asserting that the r.vs. X and Y are independent r.vs. More formally, we have the following definition for independent r.vs.

Definition 5 : Let X and Y be two r.vs. with joint p.m.f. $f(x_j, y_k)$ and marginal p.m.fs. $g(x_j)$ and $h(y_k)$ of X and Y, respectively. If for **all** pairs (x_j, y_k),

$$f(x_j, y_k) = g(x_j)\, h(y_k), \qquad \ldots (3)$$

then we say that the r.vs. X and Y are **stochastically independent** or, simply, **independent.**

Now we give an equivalent definition in the following remark.

Remark 1 : An equivalent definition of independence of X and Y would be as follows : The r.vs. X and Y are independent if for all pairs (x_j, y_k), the events $[X = x_j]$ and $[Y = y_k]$ are independent, i.e., if

$$P[X = x_j, Y = y_k] = P[X = x_j]\, P[Y = y_k] \qquad \ldots (4)$$

for **all** pairs (x_j, y_k).

Note that we have defined independence of r.vs. in terms of independence of events. Thus, no essentially new concept is involved in the definition of independence of two r.vs. except that the product relation (3) or equivalently the product relation (4) should hold for **all** pairs (x, y) of values x of X and y of Y.

Note that the r.vs. X and Y of Examples 7 and 8 are **not** independent. You can check that

$$f(0, 0) \neq g(0)\, h(0) \text{ in Example 7.}$$

Similarly, in Example 8,

$$f(1, 2) \neq g(1)\, h(2).$$

With this background, can you extend the concept of independence of two r.vs. to that of n(> 2) r.vs?

Definition 6 : Let $X_1, \ldots, X_n$ be n r.vs. They are said to be **independent** if

$$P[X_1 = x_1, \ldots, X_n = x_n] = P[X_1 = x_1]\, P[X_2 = x_2] \ldots P[X_n = x_n]$$

for **all** n-tuples $(x_1, \ldots, x_n)$ of values x_1 of X_1, x_2 of X_2, ..., x_n of X_n.

If you have followed the ideas introduced in this section, then you should be able to solve these exercises.

E7) Determine the value of c so that the following functions represent the joint p.m.f. of the r.vs. X and Y.

a) $f(x, y) = c$, $x = 1, 2, 3$, $y = 1, 2, 3$.

b) $f(x, y) = c(x^2 + y^2)$, $x = -1, 1$, $y = -2, 2$.

c) $f(x, y) = c(x + y + 1)$, $x = 0, 1, 2, 3$, $y = 0, 1, 2$.

E8) Obtain the marginal p.m.fs. of X and Y in each of the cases of E7.

...independent in each of the cases of E7.

E10) Suppose the r.vs. X and Y have the joint p.m.f. f(x, y) specified by the following table

X \ Y	0	1
1	0.20	0.15
2	0.20	0.30
3	0.05	0.10

a) Obtain the marginal p.m.fs. of X and Y.

b) Determine if X and Y are independent.

So far, you have seen that the p.m.f. of one or more random variables can be visualised as their frequency distribution where probabilities correspond to relative frequencies. You also know that given a frequency distribution, we can find its mean, variance, covariance and moments. Let us study these concepts for the p.m.f. of a r.v. now.

7.4 MATHEMATICAL EXPECTATION

Suppose that the scores obtained by five students in a class are

40, 50, 55, 60 and 75.

What is the average or arithmetic mean score of these five students ? This average is

$$\frac{(40 + 50 + 55 + 60 + 75)}{5} = 56.00$$

The problem becomes a little more complicated if we have the following frequency distribution of the scores of 100 students in the class.

Score	40	50	55	60	75
Frequency	10	15	35	25	15

By the usual formula you can compute the average score as

$$\frac{1}{100}\{10 \times 40 + 15 \times 50 + 35 \times 55 + 25 \times 60 + 15 \times 75\}$$

$$= 57.00$$

However, let us rewrite this in a slightly different form as follows. The required average is

$$40 \times \frac{10}{100} + 50 \times \frac{15}{100} + 55 \times \frac{35}{100} + 60 \times \frac{25}{100} + 75 \times \frac{15}{100}.$$

Note that the fraction 10/100, 15/100, 35/100, 25/100 and 15/100 are, in fact, the relative frequencies or the proportions of the students who obtain the scores 40, 50, 55, 60 and 75, respectively.

As you know, the arithmetic mean is a measure of central tendency giving a single number around which the observations are distributed. Now we want to define a similar measure of central tendency for the probability distributions of a r.v. X, which assumes different values with their associated probabilities. The only difference is that the role of relative frequencies is now taken over by the probabilities.

The simplest situation is to consider a r.v. X which takes two values 1 and 2, and suppose that P[X = 1] = 1/3 and P[X = 2] = 2/3. The mean, or the mathematical expectation, of this r.v. X is defined to be

$$1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} = \frac{5}{3}.$$

Note that 5/3 is an average value and is not necessarily a possible value of the r.v. X (see Example 9).

Suppose now that a r.v. X takes a finite number n of values $x_1, x_2, \ldots, x_n$ with probabilities $f(x_1), f(x_2), \ldots, f(x_n)$.

Then we define the expectation of X as

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_n f(x_n)$$

$$= \sum_{j=1}^{n} x_j f(x_j) \qquad \ldots (5)$$

Suppose now that the r.v. X assumes an infinity of values $x_1, x_2, \ldots$ with associated probabilities $f(x_1), f(x_2), \ldots$ The expectation of X is now defined by the infinite series.

The symbol E(X) is read as the expectation of X.

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots$$

$$= \sum_{j=1}^{\infty} x_j f(x_j) \qquad \ldots (6)$$

provided the infinite series converges absolutely, i.e., provided $\sum_{j=1}^{\infty} | x_j | f(x_j)$ is a convergent series.

Notice that if $\sum_j | x_j | f(x_j)$ is a convergent series, then

$$| E(X) | = | \sum_j x_j f(x_j) |$$

$$\leq \sum_j | x_j | f(x_j) < \infty,$$

i.e. E(X) is a finite number or we say that E(X) is finite.

A series $\sum_j a_j$ is called a **convergent series** if $S_n = \sum_{i=1}^{n} a_i$ tends to a finite limit as $n \to 0$. But do not spend much time over this definition. You will be asked to sum only geometric series in this course.

Formally, we have the following definition which is valid both when X assumes a finite number of values and when it assumes a countably infinite number of values.

Definition 7 : The **expectation E(X)** of the r.v.X assuming values $x_1, x_2, \ldots$ with probabilities $f(x_1), f(x_2), \ldots$ is given by

$$E(X) = \sum_j x_j P[X = x_j] = \sum_j x_j f(x_j),$$

provided $\sum_j | x_j | f(x_j)$ is finite.

We shall not discuss the definition of E(X) when the infinite series $\sum | x_j | f(x_j)$ does not converge. The discussion of such cases is beyond the scope of this course and so, we shall consider only those r.vs. which have a finite expectation.

The mean of X, expected value of X, mathematical expectation of X, mean of the distribution of X are some of the synonyms in use for E(X).

We now illustrate the computation of E(X) through some examples.

Example 9 : Let us find the expected score obtained on the roll of an unbiased die.

The score X obtained on the roll of a die is 1, 2, 3, 4, 5 or 6 and each has probability 1/6, i.e. P[X = x] = 1/6 for x = 1, 2, . . . , 6. Hence,

$$E(X) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6$$

$$= 3.5$$

Example 10 : A lottery consists of 100 tickets valued at Rs. 2/ each. A person buys 1 ticket and would gain a prize of Rs. 100 if his ticket is the winning ticket. Let us find his expected gain if the winning ticket is selected at random out of the 100 tickets.

The probability that the person wins the prize is 1/100 and that he loses is 99/100. His net gain X is Rs. 98 if he wins, and is Rs. (–2) if he loses. Thus, we need to find E(X) when P[X = –2] = 99/100 and P[X = 98] = 1/100. We get

$$E(X) = 98 \times \frac{1}{100} + (-2) \times \frac{99}{100} = -1.$$

Thus his net expected gain is Rs. (-1), i.e., his expected loss is Rs. 1.

Now we consider two situations, where the r.v. takes an infinite number of values.

Example 11 : Suppose we want to find the expected value of a r.v. X which has the p.m.f.

$$f(x) = \frac{2}{3}(1/3)^x,\ x = 0, 1, 2 \quad \ldots$$

By definition

$$E(X) = \sum_{x=0}^{\infty} x\, f(x) = \sum_{x=0}^{\infty} \frac{2}{3}\left(\frac{1}{3}\right)^x$$

$$= \frac{2}{3} \sum_{0}^{\infty} \left(\frac{1}{3}\right)^x$$

$$= \frac{1}{2}.$$

Many a times, we need to calculate not E(X) but the expected value of a function of X, like X^2, cos X, exp(tX), etc. Of course, all such functions are again r.vs. and we can use the definition to calculate their expectation. However, the following example suggests a simple solution.

Example 12 : Let X be a r.v. with p.m.f. given by the following table.

x	–2	–1	0	1	2
f(x)	1/10	2/10	4/10	2/10	1/0

We want to compute $E(X^2)$.

Since X assumes the values –2, –1, 0, 1, 2, the values of X^2 are 0, 1 and 4. Do you agree that

$$P[X^2 = 0] = P[X = 0] = 4/10\ ?$$

Now, since $X^2 = 1$ iff X = 1 or X = –1,

$$P[X^2 = 1] = P[[X = 1] \cup [X = 1]]$$

$$= P[X = 1] + P[X = -1] = 4/10.$$

Similarly, $P[X^2 = 4] = P[X = 2] + P[X = -2] = 2/10$.

In short, the p.m.f. of X^2 is specified by

$$P[X^2 = 0] = 4/10,\ P[X^2 = 1] = 4/10,\ P[X^2 = 4] = 2/10.$$

Hence,

$$E(X^2) = 0 \times \frac{4}{10} + 1 \times \frac{4}{10} + 4 \times \frac{2}{10}$$

$$= 1.2 \qquad \ldots (7)$$

Here we first obtained the p.m.f. of X^2 and then used the definition of $E(X^2)$. This, in general, could be a cumbersome procedure. So let's try another way.

Let us calculate $\sum_{x=-2}^{2} x^2 f(x)$.

$$\sum_{x=-2}^{2} x^2 f(x) = 4\left(\frac{1}{10}\right) + 1\left(\frac{2}{10}\right) + 0\left(\frac{4}{10}\right) + 1\left(\frac{2}{10}\right) + 4\left(\frac{1}{10}\right)$$

$$= 1.2 \qquad \ldots (8)$$

The equality $E(X^2) = \sum_{x=-2}^{2} x^2 f(x)$, brought out by (7) and (8) is not an accident. It is a consequence of some detailed analysis which leads us to the following theorem.

Theorem 1 : Let X be a r v assuming values $x_1, x_2, \ldots$ with probabilities $f(x_1), f(x_2), \ldots$ Let $\phi(X)$ be a r.v. which is a function of X, i.e., when $X = x_j$, $\phi(X) = \phi(x_j)$. Then

$$E[\phi(X)] = \sum_j \phi(x_j)\, f(x_j), \qquad \ldots (9)$$

provided the series on the right hand side of (9) is absolutely convergent.

We shall not prove this theorem. But we would like to bring out some important points concerning it.

Remark 2 :

i) We have the following useful interpretation for $E[\phi(x)]$:

$$E[\phi(X)] = \sum_j \phi(x_j)\, P[X = x_j].$$

ii) The illustration in Example 12 is **not** a proof of the above theorem. The proof is beyond the scope of this course.

We'll be interested in functions of the type $\phi(x_j, y_k) = x_j + y_k$, $\phi(x_j, y_k) = x_j$, $\phi(x_j, y_k) = x_j y_k$.

iii) Suppose X and Y are two r.vs. with joint p.m.f. $f(x_j, y_k)$. Let ϕ be a real-valued function defined on the product set $G \times H$, where $G = \{x_1, x_2, \ldots\}$ is the set of values of X and $H = \{y_1, y_2, \ldots\}$ is the set of values of Y.

Let us denote by $\phi(X, Y)$, the r.v. which assumes the value $\phi(x_j, y_k)$, when $X = x_j$ and $Y = y_k$. We define, by analogy with the result of Theorem 1,

$$E[\phi(X, Y)] = \sum_j \sum_k \phi(x_j, y_k)\, f(x_j, y_k), \qquad \ldots (10)$$

provided, of course, the infinite series on the right is absolutely convergent.

Now consider the random variable $\phi(X) = aX + b$, where a and b are constants. What will be then the expectation of $aX + b$? Suppose X assumes the values $x_1, x_2, \ldots$ with probabilities $f(x_1), f(x_2), \ldots$ We have

$$E[aX + b] = \sum_j (ax_j + b)\, f(x_j) \quad \text{by (9)}$$

$$= a \sum_j x_j f(x_j) + b \sum_j f(x_j)$$

$$= aE(X) + b, \text{ since } \sum x_j f(x_j) = E(X) \text{ and } \sum f(x_j) = 1.$$

We can generalise this result and find a simple way of calculating the expectation of the sum of two r.vs. X and Y. This is given in the following result.

Suppose X and Y are two r.vs. with joint p.m.f., $f(x_j, y_k)$, $j, k = 1, 2, \ldots$ Suppose E(X) and E(Y) are finite. Then E(X + Y) is finite and

$$E(X + Y) = E(X) + E(Y).$$

This result is true when X and Y take either finite or countably infinite values. We shall not worry about the proof in the countably infinite case here. The proof in the finite case is very easy and we are sure you can write it yourself.

E11) If X and Y are two r.vs. with joint p.m.f. $f(x_j, y_k)$, j = 1, 2, . . . , n, k = 1, 2, . . . m, and E(X), E(Y) are finite, then prove that

$E(X + Y) = E(X) + E(Y)$.

A simple induction argument leads us to the following result:

If $X_1, X_2, \ldots, X_n$ are r.vs. such that $E(X_i)$ is finite for all i, then $X_1 + \ldots + X_n$ also has a finite expectation and

$$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + \ldots + E(X_n).$$

We now list a simple but useful property of E(X).

If $a \leq X \leq b$, a, b . **R**, i.e., if the values $x_1, x_2, \ldots$ of the r.v. X are such that $a \leq x_j \leq b$ for all j = 1, 2, . . . , then a E(X) b.

Proof : Observe that because $a \leq x_j \leq b$ for all $j \geq 1$, we have

$$a \sum_j f(x_j) \leq \sum_j x_j f(x_j) \leq b \sum_j f(x_j).$$

Equivalently, since $\sum f(x_j) = 1$, $a \leq E(X) \leq b$.

See if you can solve these exercises by using the results of this section.

E12) Prove :

a) If $X \geq 0$, and E(X) is finite than E(X) *symbol* 0.

b) Let $X \geq Y$, i.e., the r.v. X – Y assumes only non-negative values. Then

$E(X) \geq E(Y)$.

E13) Let the p.m.f. of a r.v. X be

$f(x) = (3 - x)/10$, $x = -1, 0, 1, 2$.

a) Calculate E(X).

b) Calculate $E(X^2)$ by using (9) and also by determining the p.m.f. of X^2 and verify that both give the same result.

c) Use the results of (a) and (b) to calculate

$E[(4X + 5)^2]$.

E14) Calculate E[exp(tX)] for the distribution discussed in Example 11. Here t is a fixed number.

E15) An unbiased die is rolled. We say that a success occurs if the score obtained is 1 or 2. Any other score (i.e. a score of 3, 4, 5 or 6) is called a failure. Let $X_k = 0$ or 1 according as the k-th trial results in a failure or a success. Notice that $X_1 + \ldots + X_n$ is the number of successes obtained in n rolls of the die. Obtain $E(X_k)$ and hence the expected number of successes in n rolls of the die.

So far we have discussed some of the properties of the expectation of a r.v. X. You have seen that expectation is regarded as a measure of central tendency of the probability distribution of X, with the probabilities $f(x_j) = P[X = x_j]$ playing the role of relative frequencies. In the next section we will extend these concepts to obtain measures of dispersion of X around its mean value.

7.5 VARIANCE, COVARIANCE AND CORRELATION COEFFICIENT

Now we will talk about measures of dispersion of X around its mean value. We shall also introduce measures of correlation between two r.v.s. These measures are similar to the

measures of dispersion and correlation you studied in Block 1. In what follows, we assume that all the relevant expectations are defined.

Definition 8 : Let X be a r.v. assuming values $x_1, x_2 \ldots$ with probabilities $f(x_1), f(x_2) \ldots$ Let μ denote E(X). The variance of X, denoted by Var (X), is

$$\text{Var}(X) = E[(X-\mu)^2] = \sum_j (x_j - \mu)^2 f(x_j) \qquad \ldots (15)$$

Note that as we have seen in the case of the expectation, Var (X) has a close similarity with the variance (or the second moment about the mean) of a frequency distribution discussed in Block 1.

The expression (15) for Var (X) is not suitable for purposes of computation. The following lemma provides a simplification.

Lemma 1 : $\text{Var}(X) = E(X^2) - \mu^2$. ...(16)

Recall that we have proved a similar result in Sec. 2.4.3 of Block 1.

The proof of this lemma follows on exactly similar lines.

It is also convenient to write (16) as

$$\text{Var}(X) = E\{(X-\mu)^2\} = E(X^2) - [E((X)]^2 \qquad \ldots (17)$$

The positive square root of Var(X) is called the **standard deviation** of X. We denote it by $\sigma(X)$.

The variance of X, being the expectation of the non-negative r.v. $(X-\mu)^2$, is always non-negative, i.e. $\text{Var}(X) \geq 0$ (see E12 a). Also Var(X) is finite, whenever $E(X^2)$ is finite (see E12 a).
For, suppose $E(X^2)$ is finite. Then since $|X| \leq X^2 + 1$, $E[|X|] \leq E(X^2) + 1$, and hence $E[|X|]$ is finite. So, whenever $E(X^2) < \infty$, $E[|X|]$ is finite and so, be definition, E(X) is finite. Then (17) implies that Var(X) is finite.

Note further that if X is a r.v. such that P[X = a] = 1, then E(X) = a. It also follows that P[X – a = 0] = 1, implying $E[(X-a)^2] = 0$. Hence, if the r.v. X assumes only one value, its variance is zero. Conversely, if Var(X) = 0,

$$\sum_j (x_j - \mu)^2 f(x_j) = 0,$$

This implies that $(x_j - \mu)^2 = 0$ for all j such that $f(x_j) > 0$. This means that X takes only one value μ, or that $P[X = \mu] = 1$. In short, Var (X) is zero iff the r.v. X assumes only one value or is a constant. Such a r.v. is said to have a **degenerate distribution** or is said to be a **degenerate r.v.**

Now look at some examples, where we have calculated the variance or some r.vs., which you have already met.

Example 13 : Here we calculate the variance of the score obtained on the throw of an unbiased die.

Let X denote the score obtained on the throw of the unbiased die. Then

$$f(x) = P[X = x] = \frac{1}{6}, \quad x = 1, 2, \ldots, 6.$$

In Example 9 we have seen that

$$E(X) = 3.5.$$

Further, $E(X^2) = \frac{1}{6}\{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2\}$

$$= \frac{91}{6}.$$

Hence, $\text{Var}(X) = \frac{91}{6} - (3.5)^2 = \frac{35}{12}$.

Example 14: Let us calculate the variance of the gain of the person of Example 10.

Recall that $P[X = -2] = \frac{99}{100}$ and $P[X = 98] = \frac{1}{100}$ and that $E(X) = -1$. Hence,

$$\text{Var}(X) = (-2)^2 \cdot \frac{99}{100} + (98)^2 \cdot \frac{1}{100} - (-1)^2$$

$$= 99.$$

Example 15 : Suppose we want to obtain the variance of the r.v. X of Example 11.

Since $P[X = x] = \frac{2}{3}\left(\frac{1}{3}\right)^x$, $x = 0, 1, 2, \ldots$ we have

$$E(X^2) = \frac{2}{3}\sum_{x=0}^{\infty} x^2 \left(\frac{1}{3}\right)^x$$

$$= \frac{2}{3}\left\{1^2 \cdot \frac{1}{3} + 2^2 \cdot \left(\frac{1}{3}\right)^2 + 3^2 \cdot \left(\frac{1}{3}\right)^3 + \ldots\right\}$$

$$= 1.$$

Thus, $\text{Var}(X) = 1 - \left(\frac{1}{2}\right)^2 = 0.75.$

In Example 11 as well as in the above example, we were required to calculate the sums of some infinite series. Here is how we find the sums of series of the type,

$$S_0 = \sum_{j=1}^{\infty} p^j, \quad S_1 = \sum_{j=1}^{\infty} j\,p^j, \quad S_2 = \sum_{j=1}^{\infty} j^2 p^j, \; 0 < p < 1.$$

Using the formula for the sum of a geometric series, we get

$$S_0 = \frac{p}{1-p}.$$

To compute S_1, note that

$$(1-p)S_1 = \sum_{j=1}^{\infty} j\,p^j - \sum_{j=1}^{\infty} j\,p^{j+1} = \sum_{j=1}^{\infty} \{j - (j-1)\}\,p^j$$

$$= S_0 = \frac{p}{1-p},$$

Therefore, $S_1 = \frac{p}{(1-p)^2}$

Similarly,

$$(1-p)\,S_2 = \sum_{j=1}^{\infty} \{j^2 - (j-1)^2\}\,p^j$$

$$= \sum_{j=1}^{\infty} \{2j-1\}\,p^j$$

$$= 2S_1 - S_0.$$

This gives us

$$S_2 = \frac{p(1+p)}{(1-p)^3}$$

The calculations in Example 15 are for $p = 1/3$.

Now here is an exercise for you.

E16) Prove that $Var(aX + b) = a^2 Var(X)$.

We now give some important observations concerning the result in E16 in the following remark.

Remark 3 :

i) If we treat Y as a r.v. obtained from X by a change of origin and scale, then E16 implies that the variance is unaffected by change of origin.

ii) The standard deviation of $Y = aX + b$ is $|a|$ times the standard deviation of X.

iii) Suppose $E(X) = \mu$ and $Var(X) = \sigma^2$, where σ is the standard deviation of X.

Then the mean and variance of $Y = \frac{(X - \mu)}{\sigma}$, are zero and one, respectively. The r.v. $Y = \frac{(X - \mu)}{\sigma}$, is called the **standardized** or **normalized version of X.**

Our next aim is to obtain Var(X+Y). For this purpose we need to introduce the concept of covariance of the two r.vs. X and Y.

Let X and Y be two r.vs. with joint p.m.f. $f(x_j, y_k)$, $j, k = 1, 2, \ldots$ Then

$$E(XY) = \sum_j \sum_k x_j y_k f(x_j, y_k), \qquad \ldots (19)$$

$(x_j \pm y_k)^2 \geq 0$
$\Rightarrow \frac{x_j^2 + y_k^2}{2} \geq |x_j y_k|$

where the sum of the series on the right is assumed to be finite (see remark 2(iii)). Let μ_x and μ_y denote the means of X and Y, respectively. Now we are in a position to define the covariance.

Definition 9 : The **covariance between X and Y**, is defined to be

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

$$= \sum_j \sum_k (x_j - \mu_x)(y_k - \mu_y) f(x_j, y_k) \qquad \ldots (20)$$

We can simplify this as follows :

$$\begin{aligned} Cov(X, Y) &= E\{XY - \mu_x Y - \mu_y X + \mu_x \mu_y\} \\ &= E(XY) - \mu_x E(Y) - \mu_y E(X) + \mu_x \mu_y \\ &= E(XY) - \mu_x \mu_y, \qquad \ldots (21) \end{aligned}$$

We can also write

$$Cov(X, Y) = E(XY) - E(X) E(Y). \qquad \ldots (22)$$

The elementary inequality $|x_j y_k| \leq \frac{(x_j^2 + y_k^2)}{2}$ implies that

$$\sum_j \sum_k |x_j y_k| f(x_j, y_k) \leq \frac{1}{2} \sum_j \sum_k (x_j^2 + y_k^2) f(x_j, y_k).$$

Hence, we conclude that

$$E[|XY|] \leq E\left(\frac{X^2 + Y^2}{2}\right) = \frac{1}{2}[E(X^2) + E(Y^2)].$$

It follows that if Var(X) and Var(Y) are finite, then Cov(X, Y) is finite.

We illustrate the procedure for the computation of Cov(X, Y) by means of an example now.

Example 16 : Suppose the joint p.m.f. of X, Y is given by the following table :

Table 5

x \ y	0	1	2	g(x)
0	3/28	9/28	3/28	15/28
1	3/14	3/14	0	3/7
2	1/28	0	0	1/28
h(y)	5/14	15/28	3/28	1

Let's compute the covariance Cov (X, Y).

We have

$$\mu_x = E(X) = \sum_{x=0}^{2} x g(x)$$

$$= 0 \times \frac{15}{28} + 1 \times \frac{3}{7} + 2 \times \frac{1}{28}$$

$$= \frac{1}{2}.$$

Similarly,

$$\mu_y = E(Y) = \sum_{y=0}^{2} y h(y)$$

$$= 0 \times \frac{5}{14} + 1 \times \frac{15}{28} + 2 \times \frac{3}{28}$$

$$= \frac{3}{4}.$$

Moreover,

$$E(XY) = 0 \times 0 \times \frac{3}{28} + 0 \times 1 \times \frac{9}{28} + 0 \times 2 \times \frac{3}{28} + 1 \times 0 \times \frac{3}{14}$$

$$+ 1 \times 1 \times \frac{3}{14} + 1 \times 2 \times 0 + 2 \times 0 \times \frac{1}{28} + 2 \times 1 \times 0 + 2 \times 2 \times 0$$

$$= \frac{3}{14}.$$

Hence, Cov (X, Y) = E (XY) – $\mu_x \mu_y$

$$= \frac{3}{14} - \frac{1}{2} \cdot \frac{3}{4}$$

$$= -\frac{9}{56}.$$

You must have noticed that the troublesome step in this calculation is the computation of E(XY). But for some r.vs., this is simplified. We establish this in the following theorem.

Theorem 2 : If X and Y are independent r.vs. and have finite expectations, then

$$E(XY) = E(X)\, E(Y).$$

Proof : Since X and Y are independent r.vs., their joint p.m.f. is

$$f(x_j, y_k) = g(x_j)\, h(y_k),$$

Where $g(x_j)$ and $h(y_k)$ are the marginal p.m.fs. of X and Y, respectively (see Definition 5).

We, therefore, have

$$E(XY) = \sum_j \sum_k x_j\, y_k\, f(x_j, y_k)$$

$$= \sum_i \sum_k x_j\, y_k\, g(x_j)\, h(y_k)$$

$$= \left\{\sum_j x_j\, g(x_j)\right\}\left\{\sum_k y_k\, h(y_k)\right\}$$

$$= E(X)\, E(Y).$$

We generalise this result for n independent r.vs in the following corollary.

Corollary : If $X_1, X_2, \ldots X_n$, are n independent r.vs with finite expectations, then

$$E\left(\prod_{j=1}^{n} X_j\right) = E(X_1)\;\; E(X_2) \ldots . E(X_n).$$

We are not going to prove this corollary here.

Here is another useful result which follows from Theorem 2:

Corollary: If X and Y are independent r.vs. with finite variances, then Cov (X,Y) = 0.

Caution : If Cov(X, Y) = 0, it does not follow that X and Y are independent. For example, consider the r.vs. X and Y with joint p.m.f. as in Table 6.

Table 6

x \ y	0	1	2	3	g(x)
1	2/27	0	0	1/27	3/27
2	6/27	6/27	6/27	0	18/27
3	0	6/27	0	0	6/27
h(y)	8/27	12/27	6/27	1/27	1

Observe that

$$E(X) = \frac{3}{27} + \frac{36}{27} + \frac{18}{27} = \frac{19}{9},$$

$$E(Y) = \frac{12}{27} + \frac{12}{27} + \frac{3}{27} = 1,$$

and

$$E(XY) = 1 \cdot 3 \cdot \frac{1}{27} + 2 \cdot 1 \cdot \frac{6}{27} + 2 \cdot 2 + \frac{6}{27} + 3 \cdot 1 \cdot \frac{6}{27}$$

$$= \frac{19}{9}.$$

Thus, Cov (X, Y) = 0.

However, $f(1, 1) = 0 \neq g(1)\, h(1)$. This shows that X and Y are not independent.

This, X and Y are independent $\Rightarrow$ X and Y have zero covariance.

But the converse is **not true.**

We are now in a position to obtain Var (X + Y).

If X and Y are random variables with finite variances, then

$$\text{Var}\,(X + Y) = \text{Var}\,(X) + \text{Var}\,(Y) + 2\text{Cov}\,(X,Y).$$

Let's prove this

Proof : Let X and Y be r.vs. with joint p.m.f. $f(x_j, y_k)$. Then $E(X + Y) = \mu_x + \mu_y$ and

$$\text{Var}\,(X + Y) = E\left[\{X + Y - \mu_x - \mu_y\}^2\right]$$

$$= \sum_j \sum_k \{x_j + y_k - \mu_x - \mu_y\}^2 f(x_j, y_k)$$

$$= \sum_j \sum_k \{x_j - \mu_x + y_k - \mu_y\}^2 f(x_j, y_k)$$

$$= \sum_j \sum_k (x_j - \mu_x)^2 f(x_j, y_k) + \sum_j \sum_k (y_k - \mu_y)^2 f(x_j, y_k)$$

$$+ 2 \sum_j \sum_k (x_j - \mu_x)(y_k - \mu_y) f(x_j, y_k)$$

$$= \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$$

as required.

Corollary: If X and Y are r.vs. with Cov (X, Y) = 0, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \qquad \ldots (23)$$

Note that if X and Y are independent r.vs., then (23) automatically holds.

We now give a result about the variance of the sum of n r.vs :

If $X_1, X_2, \ldots, X_n$ are r.vs with finite variances,

$$\text{Var}(X_1 + \ldots + X_n) = \sum_{j=1}^{n} \text{Var}(X_j) + 2 \sum_{j=1}^{n} \sum_{k=j+1}^{n} \text{Cov}(X_j, X_k).$$

We omit the proof of this result. The result about n independent r.vs now follows:

Corollary : if $X_1, \ldots, X_n$ are independent r.vs., with finite variances, then

$$\text{Var}(X_1 + \ldots + X_n) = \sum_{j=1}^{n} \text{Var}(X_j).$$

In fact, it is enough to assume that the r.vs. $X_1, \ldots, X_n$ have pairwise zero covariances to claim this result. Try to do this exercise now. It concerns the definitions and results which we have just discussed.

E17) Let the joint distribution of X and Y be as specified in Example 16. Obtain Var (X + Y).

The following theorem expresses the covariance between aX + b and cY + d, where a, b, c, d are constants, in terms of Cov(X, Y).

Theorem 4 : Cov(aX + b, cY + d) = ac Cov (X, Y).

Proof : Let $Z_1 = aX + b$ and $Z_2 = cY + d$ Then $E(Z_1) = a\mu_x + b$ and $E(Z_2) = c\mu_y + d$. Now

$$\text{Cov}(aX + b, cY + d) = E(Z_1 Z_2) - E(Z_1)E(Z_2)$$

$$= E[(aX + b)(cY + d)] - (a\mu_x + b)(c\mu_y + d)$$

$$= E[acXY + adX + bcY + bd]$$

$$- ac\mu_x\mu_y - ad\mu_x - bc\mu_y - bd$$

$$= ac\,E(XY) + ad\mu_x + bc\mu_y + bd - ac\mu_x\mu_y$$

$$- ad\mu_x - bc\mu_y - bd$$

$$= ac\{E(XY) - \mu_x\mu_y\}$$

$$= ac\,\text{Cov}(X, Y), \text{ as required.}$$

We can use this theorem to arrive at the following result.

Corollary : If X and Y are r.vs. with Cov(X, Y) = 0, then

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y).$$

Proof : Applying Theorem 4, we get

$$\text{Cov}(X, -Y) = -\text{Cov}(X, Y)$$

Also, be using the result in E16, we can write Var (–Y) = Var (Y).

Hence, in general,

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(-Y) + 2\,\text{Cov}(X, -Y).$$

$$= \text{Var}(X) + \text{Var}(Y) - 2\,\text{Cov}(X, Y).$$

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y).$$

We conclude this section with the discussion of the correlation coefficient between X and Y. The definition of the correlation coefficient is very similar to that of the correlation coefficient you encountered in Unit 4 in connection with bivariate data.

Definition 10 : The **correlation coefficient** between X and Y is defined to be

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}.$$

In this definition we assume that Var (X) and Var (Y) are both finite and positive and that the square root in the denominator is the positive square root. We give below some simple properties of the correlation coefficient.

1) Let $Z_1 = aX + b$ and $Z_2 = cY + d$. Then

$$\rho(Z_1, Z_2) = \begin{cases} \rho(X, Y) \text{ if } ac > 0, \\ -\rho(X, Y) \text{ if } ac < 0. \end{cases}$$

Proof : Recall that $\text{Cov}(Z_1, Z_2) = ac\,\text{Cov}(X, Y)$ and that $\text{Var}(Z_1) = a^2\,\text{Var}(X)$ and $\text{Var}(Z_2) = c^2\,\text{Var}(Y)$. Hence

$$\rho(Z_1, Z_2) = \frac{(ac)\,\rho(X, Y)}{|ac|},$$

from which the required result follows.

In particular, if $X^* = \frac{(X - \mu_x)}{\sigma_x}$ and $Y^* = \frac{(Y - \mu_y)}{\sigma_y}$ are the standardised versions of X and Y, respectively, then $\rho(X^*, Y^*) = \rho(X, Y)$, since σ_x and σ_y are both positive.

2) $-1 \le \rho(X, Y) \le +1$.

Proof : You have already seen this result in Unit 4. Here is an alternative proof. Let X* and Y* be the standardised r.vs. Since $\text{Var}(X^*) = \text{Var}(Y^*) = 1$, we find that $\text{Cov}(X^*, Y^*) = \rho(X^*, Y^*) = \rho(X, Y)$. Moreover,

$$\text{Var}(X^* + Y^*) = \text{Var}(X^*) + \text{Var}(Y^*) + 2\text{Cov}(X^*, Y^*)$$

$$= 2\{1 + \rho(X, Y)\}$$

Since $\text{Var}(X^* + Y^*) \ge 0$, we have

$$2[1 + \rho(X, Y)] \ge 0,$$

or $\rho(X, Y) \ge -1$.

Similarly

$$0 \le \text{Var}(X^* - Y^*) = 2\{1 - \rho(X, Y)\}$$

implies that $\rho(X, Y) \le 1$. Hence the result.

3) The correlation coefficient $\rho(X, Y) = \pm 1$ if and only if there exist connstants a and b such that $Y = aX + b$.

Proof : Let $Y = aX + b$. Then

$$\text{Var}(Y) = a^2\,\text{Var}(X) \text{ and}$$

$$\text{Cov}(X, Y) = \text{Cov}(X, aX + b) = a\,\text{Cov}(X, X) = a\,\text{Var}(X).$$

Hence $\rho(X, Y) = \frac{a\text{Var}(X)}{\sqrt{\{a\,\text{Var}(X)\}^2}}$

$$= \pm 1 \text{ according as } a > 0 \text{ or } a < 0.$$

Conversely, suppose $\rho(X, Y) = 1$. Then from the proof of the second property above, we have

$$\text{Var}(X^* - Y^*) = 0.$$

This implies that $X^* - Y^*$ is a degenerate r.v. or that

$$X^* - Y^* = c, \text{ a constant}$$

Since $E(X^*) = E(Y^*) = 0$, $c = 0$. Equivalently,

$$\frac{X - \mu_x}{\sigma_x} - \frac{Y - \mu_y}{\sigma_y} = 0$$

$$\text{or } Y = \frac{\sigma_y}{\sigma_x} X + \mu_y - \frac{\sigma_y}{\sigma_x} \mu_x$$

$$= aX + b,$$

where $a = \frac{\sigma_y}{\sigma_x}$ and $b = \mu_y - \frac{\sigma_y}{\sigma_x} \mu_x$.

The proof for the case when $\rho(X, Y) = -1$ is similar. In that case we use the result

$$\text{Var}(X^* + Y^*) = 0.$$

We have given a number of examples in this section to show how to obtain the mean, variance, covariance, etc. for random variables. Now would you like to try your hand at these exercises?

E18) Compute the means, the variances, the covariances and the correlation coefficients for the joint distribution of E7 and E10.

E19) Obtain the variance of the total number of successes in E15 under the assumption that $X_1, X_2, \ldots, X_n$ are independent r.vs.

E20) Obtain Var $(aX + bY)$.

So far we have discussed many concepts for a random variable with a given p.m.f. We had talked about the same concepts in relation to a frequency distribution in Block 1. In the next section we will take up the study of yet another concept.

7.6 MOMENTS AND MOMENT GENERATING FUNCTION

We have studied the properties of E(X) and Var(X) in the previous two sections. There are expectations of some functions of r.vs associated with a probability distribution, which play an important role in statistical theory. We plan to study properties of some of these in this section.

Let r be a positive integer. The r-th moment of a r.v. X or of its probability distribution is

$$\mu'_r = E(X^r) = \sum_j x_j^r f(x_j),$$

provided, of course, the series on the right is absolutely convergent. Sometimes we need to use

$$m_r(a) = E[(X - a)^r] = \sum_j (x_j - a)^r f(x_j),$$

which is called the r-th moment of X about a. In this sense μ_r^1 is the r-th moment about the origin (a = 0).

Of course, when r = 0, $X^0 = 1$ and therefore, $\mu'_0 = 1$. The first moment μ'_1 is, the, by now familiar, expected value or mean of X. The variance, Var(X), is the second moment of X about its mean, $m_2(\bar{x})$.

Let u be a real number. If $|u| > 1$, then $|u|^{r-1} \leq |u|^r$ and if $|u| \leq 1$, then

$|u|^{r-1} \le 1$. Hence we can assert that whatever be the real number u, $|u|^{r-1} \le |u|^r + 1$. A consequence of this inequality is the following :

$$\sum_j |x_j|^{r-1} f(x_j) \le \sum_j \{|x_j|^r + 1\} f(x_j)$$

$$= 1 + \sum_j |x_j|^r f(x_j).$$

Thus, whenever, the r-th moment of X is finite, so is the (r – 1)-th moment. In particular, all the moments μ'_s, $s \le r$, would be finite.

We do not enter into any detailed study of the properties of moments of a rv. except to introduce the so-called moment generating function which will be useful to us in Units 8 and 9.

Let t be a real variable and suppose that

Exp (tX) = e^{tX}

$$M_x(t) = E\{\exp(tX)\} = \sum_j \exp(tx_j) f(x_j)$$

is finite for all values of t in a neighbourhood of the origin t = 0. Then the function $M_x(t)$ of t is called the **moment generating function** of X. We abbreviate it as **m.g.f.**

You may be wondering why we call $M_x(t)$, the moment generating function. Recall that Maclaurin's expansion of exp(tx) is

$$\exp(tx) = 1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \ldots \text{ (Calculus, Unit 6)}$$

It, therefore, follows that

$$M_x(t) = \sum_j \left\{1 + tx_j + \frac{t^2x_j^2}{2!} + \ldots + \frac{t^rx_j^r}{r!} + \ldots\right\} f(x_j)$$

$$= 1 + t\mu'_1 + \frac{t^2\mu'_2}{2!} + \ldots + \frac{t^r\mu'_r}{r!} + \ldots .$$

In other words, μ'_r is the coefficient of $t^r/r!$ in Maclaurin's expansion of the m.g.f. In fact, we can write

$$\mu'_r = \left[\frac{d^rM_x(t)}{dt^r}\right]_{t=0}$$

In this sense, the m.g.f. **generates** moments.

In the following example we find the m.g.f. of a random variable.

Example 17 : Let X be a r.v. with

$$P[X = 0] = 2/3 \text{ and } P[X = 1] = 1/3.$$

Its m.g.f. is

$$M_x(t) = e^{t.0} \cdot \frac{2}{3} + e^{t.1} \cdot \frac{1}{3}$$

$$= \frac{(2 + e^t)}{3}.$$

Since here

$$M_x(t) = \frac{2}{3} + \frac{1}{3}\left\{1 + t + \frac{t^2}{2!} + \ldots\right\},$$

we find that

$$\mu'_0 = \frac{2}{3} + \frac{1}{3} = 1 \text{ and } \mu'_r = \frac{1}{3}, r = 1, 2, \ldots$$

We now give some simple results about the m.g.f. which we shall use later.

I) If $Y = aX + b$, then

$$M_y(t) = e^{bt} M_x(at)$$

Proof : By definition

$$M_y(t) = E[\exp(ty)]$$

$$= E[\exp(taX + tb)]$$

$$= e^{bt} E[\exp(atX)]$$

$$= e^{bt} M_x(at).$$

In particular, if $X^* = \frac{(X - \mu_x)}{\sigma_x}$ is the standardised version of X, then

$$M_{x^*}(t) = \exp\left[\frac{-\mu_x t}{\sigma_x}\right] M_x\left(\frac{t}{\sigma_x}\right).$$

We shall use this result in some later units of this course.

The importance of the m.g.f. does not lie only in its ability to generate the moments of the r.v.X. Under certain conditions, the m.g.f. can uniquely identify the probability mass function of X and hence its probability distribution. But we'll not go into the details here.

Now we prove another result which is useful in the study of the distribution of the sum of two or more independent r.vs.

II) Let X and Y be independent r.vs with m.g.fs. $M_x(t)$ and $M_y(t)$. Then the m.g.f. of $X + Y$ is

$$M_{x+y}(t) = M_x(t)\, M_y(t).$$

Proof : Since X and Y are independent r.vs, their joint p.m.f. is $f(x_j, y_k) = g(x_j)\, h(y_k)$, where g and h are the p.m.fs. of X and Y, respectively. Hence,

$$M_{x+y}(t) = E[\exp\{t(X + Y)\}]$$

$$= \sum_j \sum_k e^{t(x_j + y_k)} f(x_j, y_k)$$

$$= \sum_j \sum_k \left\{e^{tx_j} g(x_j)\right\} \left\{e^{ty_k} h(y_k)\right\}$$

$$= \left\{\sum_j e^{tx_j} g(x_j)\right\} \left\{\sum_k e^{ty_k} h(y_k)\right\}$$

$$= M_x(t)\, M_y(t),$$

which is the required result.

We shall talk more about the probability distribution of the sum of two r.vs. in the next section. But before that we are giving you a simple exercise to do.

E21) Obtain the m.g.f. and the moments of the r.v. in Example 13.

7.7 DISTRIBUTION OF SUM OF TWO RANDOM VARIABLES

In Example 2 we have discussed the probability distribution of the sum of scores obtained on two rolls of an unbiased die. In this section we are interested in the methods of obtaining the probability distribution of the sum of two r.vs. We begin with the following simple example.

Example 18: The joint p.m.f. of (X, Y) is as specified in Table 7.

Table 7

x \ y	0	1	2	3
0	1/27	3/27	3/27	1/27
1	3/27	6/27	3/27	0
2	4/27	3/27	0	0

We want to obtain the p.m.f. of X + Y.

Observe, first of all, that since X takes the values 0, 1,2, and Y takes the values 0, 1, 2, 3, the r.v. X+Y can assume the values 0, 1, 2, 3, 4, 5. Now we list the different possibilities.

$$[X + Y = 0] = [X = 0, Y = 0],$$

$$[X + Y = 1] = [X = 0, Y = 1] \cup [X = 1, Y = 0],$$

$$[X + Y = 2] = [X = 0, Y = 2] \cup [X = 2, Y = 0] \cup [X = 1, Y = 1],$$

$$[X + Y = 3] = [X = 0, Y = 3] \cup [X = 1, Y = 2] \cup [X = 2, Y = 1],$$

$$[X + Y = 4] = [X = 1, Y = 3] \cup [X = 2, Y = 2],$$

$$[X + Y = 5] = [X = 2, Y = 3].$$

It immediately follows that

$$P[X + Y = 0] = P[X = 0, Y = 0] = 1/27$$

$$P[X + Y = 1] = P[X = 0, Y = 1] + P[X = 1, Y = 0]$$

$$= \frac{3}{27} + \frac{3}{27} = \frac{6}{27}$$

$$P[X + Y = 2] = P[X = 0, Y = 2] + P[X = 2, Y = 0] + P[X = 1, Y = 1]$$

$$= \frac{3}{27} + \frac{6}{27} + \frac{4}{27} = \frac{13}{27}$$

$$P[X + Y = 3] = P[X = 0, Y = 3] + P[X = 1, Y = 2] + P[X = 2, Y = 1]$$

$$= \frac{1}{27} + \frac{3}{27} + \frac{3}{27} = \frac{7}{27}$$

and $P[X + Y = 4] = P[X + Y = 5] = 0$.

Thus, the p.m.f. of X + Y is

$$f(0) = 1/27, f(1) = 6/27, f(2) = 13/27, f(3) = 7/27.$$

This example illustrates the general method of obtaining the p.m.f. of X+Y from the joint p.m.f. of X and Y. The basic steps are

i) Identify the possible distinct values of X + Y.

ii) If $u_1, u_2, \ldots$ denote these distinct values of X + Y, identify all the sets $[X = x_j, Y = y_k]$ for which $x_j + y_k = u_r$, say.

iii) Then

$$P[X + Y = u_r] = \sum f(x_j, y_k),$$

where the sum extends over all those (x_j, y_k) which add up to u_r.

This general procedure, though valid in principle for all discrete r.vs., is cumbersome, except in very simple situations. We, therefore, investigate a special case in which simplification is possible.

Suppose X and Y are independent r.vs. which assume non-negative integral values 0, 1, 2, . . . Let P[X = x] = f(x), and P[Y = y] = g(y), x, y = 0, 1, 2, . . . Because of the independence of X and Y,

$$P[X = x, Y = y] = f(x)\, g(y)$$

for all x and y. In order to obtain the p.m.f. of X + Y, observe that X + Y assumes the values 0, 1, 2, . . . Moreover, the event [X + Y = r] is the union of the disjoint events.

$$[X = 0, Y = r], [X = 1, Y = r - 1], \ldots, [X = r, Y = 0],$$

where r is a non-negative integer. It follows that

$$P[X + Y = r] = \sum_{j=0}^{r} P[X = j, Y = r - j]$$

$$= \sum_{j=0}^{r} f(j)\, g(r - j), \quad r = 0, 1, 2, \ldots$$

This procedure is illustrated in the following example.

Example 19: Let X and Y be independent r.vs. with

$$P[X = x] = \frac{2}{3}\left(\frac{1}{3}\right)^{x}, \quad x = 0, 1, 2, \ldots$$

$$P[Y = y] = \frac{2}{3}\left(\frac{1}{3}\right)^{y}, \quad y = 0, 1, 2, \ldots$$

i.e., X and Y are independent r.vs. with the same p.m.f. The p.m.f. of X + Y is given by

$$P[X + Y = r] = \sum_{j=0}^{r} P[X=j]P[Y=r-j]$$

$$= \left(\frac{2}{3}\right)^{2} \sum_{j=0}^{r} \left(\frac{1}{3}\right)^{j} \left(\frac{1}{3}\right)^{r-j}$$

$$= (r + 1)\left(\frac{2}{3}\right)^{2} \left(\frac{1}{3}\right)^{r}, \quad r = 0, 1, 2, \ldots$$

When you study geometric distribution in Unit 9, you will come across a more general result of which this example is a particular case.

Here are some exercises for you.

E22) Obtain the distribution of X+Y when the joint p.m.f. of X and Y is as specified in Examples 7 and 8.

This brings us to the end of this unit. In it we have discussed the probability distribution of a random variable at length. Let's now briefly recall the various concepts which we have covered here.

7.8 SUMMARY

In this unit we have covered the following points.

1) A random variable is a function defined on a sample space. Its probability distribution is specified by its p.m.f. $f(x_j) = P[X = x_j]$, j = 1, 2, . . . We can study two (or more) r.vs. X and Y in terms of their joint p.m.f., $f(x_j, y_k) = P[X = x_j, Y = y_k]$, j, k = 1, 2, . . . The marginal p.m.f., $g(x_j) = P[X = x_j]$ of X and $h(y_k) = P[Y = y_k]$ of Y can be calculated from $f(x_j, y_k)$, but the converse is not true. The r.vs. X and Y are said to be **independent** if

   $f(x_j, y_k) = g(x_j)\, h(y_k)$ for **all** pairs (x_j, y_k).

2) The expectation $E(X) = \sum_j x_j f(x_j)$ of a r.v. X with p.m.f. f(x), its variance

   $Var(X) = E(X^2) - \{E(X)\}^2$ and the covariance $Cov(X, Y) = E(XY) - E(X)E(Y)$ are

important characteristics of the r.vs. They have some simple properties like

$$E(X + Y) = E(X) + E(Y),\ E(aX) = aE(X),$$

$$\text{Var}(ax + b) = a^2 \text{Var}(X),$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y), \text{ etc.}$$

3) If X and Y are independent r.vs., they have zero covariance. But r.vs. with zero covariance are not necessarily independent. The correlation coefficient

$$\rho(X,Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\ \text{Var}(Y)}}$$

is a measure of the correlation between X and Y. It is such that $-1 \leq \rho(X, Y)$ 1, the extreme values $\rho(X, Y) = \pm 1$ being attained iff Y is a linear function of X.

4) The m.g.f. of a r.v. X is $M_x(t) = E\{\exp(tX)\}$, provided it is finite for all values of t in a neighbourhood of zero. It has the important property that if X and Y are independent, the m.g.f. $M_{x+y}(t)$ of X + Y is the product $M_x(t)\ M_y(t)$ of their m.g.fs.

5) It is possible to obtain the p.m.f. of X + Y from the joint p.m.f. of X and Y . Some simplification is possible when X and Y are non-negative integer-valued r.vs.

7.9 SOLUTIONS AND ANSWERS

E1) $c(1 + p + p^2 + \ldots) = 1, \quad 0 < p < 1.$

$$\therefore \frac{c}{1-p} = 1$$

$$\Rightarrow c = 1 - p.$$

E2) The possible values of W are –5, –4, . . . 0, 1, . . . , 5 and its p.m.f. is given by

$$f(-5) = f(5) = 1/36,\ f(-4) = f(4) = 2/36,$$

$$f(-3) = f(3) = 3/36,\ f(-2) = f(2) = 4/36,$$

$$f(-1) = f(1) = 5/36,\ f(0) = 6/36.$$

E3) The possible values of Y are 0, 1, 2, and 3. To obtain P[Y = 2] for example, observe that there are $\binom{13}{2}$ ways of selecting 2 spades out of 13 spades and the third card can be chosen in $\binom{39}{1}$ ways out of the remaining 39 cards. Hence

$$P[Y = 2] = \binom{13}{2}\binom{39}{1} \Big/ \binom{52}{3}$$

Similarly,

$$P[Y = 0] = \binom{13}{0}\binom{39}{3} \Big/ \binom{52}{3}$$

$$P[Y = 1] = \binom{13}{1}\binom{39}{2} \Big/ \binom{52}{3}$$

and

$$P[Y = 3] = \binom{13}{3}\binom{39}{0} \Big/ \binom{52}{3}$$

E4) If X = Number of attempts, the possible values of X are 1, 2, 3, 4. It is easy to check that P[X = 1] = P[X = 2] = P[X = 3] = P[X = 4] = 1/4 which gives us the p.m.f. of X. To obtain its probability distribution, we need to specify P[X . H], where H is a subset of S = {1, 2, 3, 4}. There are 16 subsets of S. In fact, we have

$$P[X \in \phi] = 0,\ P[X \in S] = 1,$$

$$P[X = 1] = P[X = 2] = P[X = 3] = P[X = 4] = 1/4$$

$$P[X \in \{1, 2\}] = P[X \in \{1, 3\}] = P[X \in \{1, 4\}]$$

$$= P[X \in \{2, 3\}] = P[X \in \{2, 4\}] = P[X \in \{3, 4\}] = 1/2$$

$P[X \in \{1, 2, 3\}] = P[X \in \{1, 2, 4\}]$

$= P[X \in \{2, 3, 4\}] = P[X \in \{1, 3, 4\}] = 3/4$

E5) (i) 2/9

(ii) 8/7

(iii) 4/9

(iv) 7/27

(v) 3/8.

E6) i) $\binom{10}{3}(1/2)^3 (1/2)^7$.

ii) $\sum_{j=4}^{10} \binom{10}{j}(1/8)^j (7/8)^{10-j}$.

iii) $\frac{10!}{3!\,2^3}\left[\frac{1}{4!\,3}(1/8)^4 (3/8)^7 + \frac{1}{5!\,2!}(1/8)^5 (3/8)^2 + \frac{1}{6!\,1!}(1/8)^6 (3/8) + \frac{1}{7!\,0!}(1/8)^7\right] \Big/ \sum_{j=4}^{10} \binom{10}{j}(1/8)^j (7/8)^{10-j}$.

E7) a) 1/9

b) 1/20

c) 1/42

E8) a) $g(x) = 1/3$, $x = 1, 2, 3$, $h(y) = 1/3$, $y = 1, 2, 3$,.

b) $g(x) = (x^2 + 4)/10$, $x = -1, 1$

$h(y) = (y^2 + 1)/10$, $y = -2, 2$

c) $g(x) = (x + 2)/14$, $x = 0, 1, 2, 3$.

$h(y) = (2y + 5)/21$, $y = 0, 1, 2$.

E9) X and Y are independent in cases a) and b).

E10) a) $g(1) = 0.35$, $g(3) = 0.50$, $g(5) = 0.15$,

$g(0) = 0.45$, $h(1) = 0.55$.

b) They are not independent.

E11) $E(X) = \sum_j x_j g(x_j) = \sum_j \sum_k x_j f(x_j, y_k)$

$E(Y) = \sum_k \sum_j y_k f(x_j, y_k)$

Since E(X) and E(Y) are finite, the series above are absolutely convergent.

$\therefore \sum_j \sum_k |x_j + y_k| f(x_j, y_k) \le \sum_j \sum_j \{|x_j| + |y_k|\} f(x_j, y_k)$

$= \sum_j \sum_k |x_j| f(x_j, y_k) + \sum_j \sum_k |y_k| f(x_j, y_k)$

$< \infty$.

$\therefore$ E (X + Y) is defined and

E (X + Y) = E(X) + E(Y).

E12) a) $X \ge 0 \Rightarrow x_j \ge \forall_j \quad = 0, 1, 2, \ldots$

$\Rightarrow E(X) = \sum x_j f(x_j)$ 0.

b) $X \geq Y \Rightarrow X - Y \geq 0 \Rightarrow E(X - Y) \geq 0$

$\Rightarrow E(X) - E(Y)) \geq 0 \Rightarrow E(X) \geq E(Y)$.

E13) a) $E(X) = 0$,

b) $E(X^2) = 1$,

c) 41.

E14) $E(e^{tx})$ is $\frac{2}{3}\{1 - e^t/3\}^{-1}$, provided $| t | < \ln 3$.

E15) $P[X_k = 0] = 4/6 = 2/3$, $P[X_k = 1] = 1/3$.

Hence $E(X_k) = 1/3$ and $E(X_1 + \ldots + X_n] = n/3$.

E16) If $Y = aX + b$, then $E(Y) = aE(X) + b$.

$$\text{Var}(aX + b) = E\left[\{y - E(Y)\}^2\right]$$

$$= E\left[\{aX - aE(X)\}^2\right]$$

$$= E\left[a^2\{X - E(X)\}^2\right]$$

$$= a^2 \text{ Var } X.$$

E17) From Example 15 we get

$$\text{Cov}(X, Y) = \frac{-9}{56}.$$

$$\text{Var}(X) = 0^2 \times \frac{15}{28} + 1^2 \times \frac{3}{7} + 2^2 \times \frac{1}{28} - \left(\frac{1}{2}\right)^2$$

$$= \frac{9}{28}.$$

$$\text{Var}((Y) = \frac{45}{112}.$$

$$\text{Var}(X + Y) = \frac{9}{28} + \frac{45}{112} + 2\left(\frac{-9}{56}\right)$$

$$= \frac{45}{112}.$$

E18) In E7), $E(X) = E(Y) = 2$,

$\text{Var}(X) = \text{Var}(Y) = 2/3$, $\text{Cov}(X, Y) = \rho(X, Y) = 0$.

$E(X) = E(Y) = 0$, $\text{Var}(X) = 1$, $\text{Var}(Y) = 4$,

$\text{Cov}(X, Y) = \rho(X, Y) = 0$.

$E(X) = 1.619$, $E(Y) = 1.191$, $\text{Var}(X) = 1.950$.

$\text{Var}(Y) = 0.630$, $\text{Cov}(X, Y) = 0.215$, $\rho(X, Y) = 0.194$.

In E10), $E(X) = 2.1$, $E(Y) = 0.55$, $\text{Var}(X) = 4.19$

$\text{Var}(Y) = 0.2475$, $\text{Cov}(X, Y) = 0.395$, $\rho(X, Y) = 0.388$.

E19) $\text{Var}(X_k) = (1/3) - (1/3)^2 = 2/9$ and hence

$\text{Var}(X_1 + \ldots + X_n) = 2n/9$.

E20) $\text{Var}(aX + bY) = a^2 \text{ Var}(X) + b^2 \text{ Var}(Y) + 2ab \text{ Cov}(X, Y)$.

E21) M.g.f. $= E[\exp(tX)]$

Now x takes values, 1, 2, . . ., 6.

M.g.f. $= \sum e^{tx_j} f(x_j)$

$$= \frac{1}{6}\left[e^t + e^{2t} + \ldots + e^{6t}\right]$$

$$= \frac{1}{6}\left[6 + 21t + \frac{91t^2}{2!} + \ldots + \frac{\left(\sum_1^6 k^6\right)}{6!} + \ldots\right]$$

$M'_0 = 1, M'_1 = \frac{7}{2}, M'_2 = \frac{91}{12}, \ldots$

E22) In Example 7, the possible values of X + Y are 0, 1, 2, 3, 4, 5 and its p.m.f. is

P[X + Y = 1] = 5/24, P[X + Y = 2] = 1/3

P[X + Y = 3] = 3/8, P[X + Y = 4] = 1/24, and

P[X + Y = 5] = 1/24; since P[X + Y = 0] = 0.

In Example 8, the possible values of X+Y are 0, 1, 2, 3, 4, 5, and its p.m.f. is

P[X + Y = 1] = 1/15, P[X + Y = 2] = 1/5,

P[X + Y = 3] = 3/10, P[X + Y = 4] = 4/15,

P[X + Y = 5] = 5/30.

UNIT 8 STANDARD PROBABILITY DISTRIBUTIONS : PART I

Structure

8.1 INTRODUCTION

In this unit and the next, we describe some frequently encountered discrete probability distributions. But what do we mean by a probability distribution? To answer this question we consider the two different situations below. Each involves repeated trials of a random experiment. At the end of these trials we have to take an appropriate decision based on the results of the experiment.

1) Suppose 25 patients of high blood pressure are given a drug and their blood pressures are measured before and after the administration of the drug. Let $u_1, \ldots, u_{25}$ and $u'_1, \ldots, u'_{25}$ denote the measurements on the corresponding patients before and after the drug is administered. Medical experts say that the drug is a success for the j-th patient if $u_j > u'_j$ and a failure if $u_j \leq u'_j$. Is it possible to decide whether the drug is effective for patients of high blood pressure on the basis of the measurements on the 25 patients included in the experiment?

2) A lady claims that she has psychic powers and that she can identify the cards drawn from a pack of 52 playing cards by any person sitting in another room. The lady is able to correctly identify 3 of the ten cards drawn with replacement from a well-shuffled pack of cards. Is her claim of psychic powers justified?

These two real life situations have one thing in common. Both involve repetitions of a random experiment a specified number of times. We have 25 repetitions in the first situation, and 10 in the second. Each trial results either in a "success" or in a "failure". Thus, success stands for $u_i > u'_i$ in the first case and, a correct guess by the lady in the second case.

The questions we asked in each case can be answered on the basis of the total number X of successes. This number X is a random variable. In Block 4 you will study the techniques which would enable you to answer the questions raised in the above illustrations. However, these techniques require that the probability distribution of the total number of successes be known.

In order to obtain this knowledge, statisticians make certain assumptions. For example, in the situations described above, we may **assume** that

i) the successive trials are independent, i.e. the outcomes of different trials are treated as independent events, and

ii) the probability of success at each trial is the same.

These assumptions would enable us to obtain the distribution of the total number of successes in n trials. We would like to emphasise here that (i) and (ii) are **assumptions made** by the statisticians, and there is no reason why the natural phenomena described above should follow these assumptions.

Nevertheless, assumptions of the above type provide us an initial solution to the problems we face.

The assumptions like those made above and the resulting probability distribution of the total number of successes are said to constitute a probabilistic or a stochastic model for our random experiment. We begin with the simplest of such models. Here we would like to point out that the models we are about to describe have been found useful in a wide variety of situations and in different disciplines like medicine, agriculture, biology, industrial engineering, psychiatry etc.

The discrete probability distributions which we shall be discussing in this unit and in Unit 9, are called standard discrete distributions, because of their wide applicability and simplicity. In this unit we'll take up the study of the Bernoulli, the binomial, the multinomial and the hypergeometric distributions. Make sure that you achieve the following objectives by the end of this unit.

Objectives

After reading this unit you should be able to:

- state the assumptions underlying the binomial, multinomial and the hypergeometric distributions,
- compute their means and variances,
- obtain the distribution of the sum of two independent binomial variates,
- compute probabilities of events associated with these standard probability distributions.

8.2 THE BERNOULLI DISTRIBUTION

We begin with the simplest probability distribution, which is the distribution of a r.v. X which assumes two values, 0 and 1. Let

$$P[X = 0] = 1 - p \text{ and } P[X = 1] = p, \qquad \ldots (1)$$

$$\text{or } P[X = x] = p^x (1 - p)^{1-x} \quad x = 0, 1,$$

where p is a number such that $0 \leq p \leq 1$. What happens when $p = 0$ or $p = 1$? When $p = 0$, $?[X = 0] = 1$, i.e. X is degenerate at zero and when $p = 1$, $P[X = 1] = 1$ i.e. X is degenerate at one. We shall usually ignore these cases.

Notice that the probability distribution of the r.v. X changes with p. Thus, in fact, (1) defines a family or a class of probability distributions of the same kind. Every member of this family is uniquely determined by the value of p and to every value of p in the interval [0, 1], there is a unique probability distribution specified by (1). It is for this reason that p is called the **parameter** of the distribution of the r.v. X.

Jacob Bernoulli (1654–1705)

The r.v. X. and its probability distribution specified by the p.m.f. (1) are, respectively, called the **Bernoulli variate** and the **Bernoulli distribution** in honour of Jacob Bernoulli (1654-1705). He made a systematic study of problems connected with this distribution.

Can you think of an example of a Bernoulli variate? What about the toss of an unbiased coin? Here $p = 1/2$. If, however, the coin is not a balanced one, p can be any value in [0, 1]. In most practical situations we would not know the value of p. Therefore, it is best to study the properties of the Bernoulli distribution for a general p. In fact, we shall adopt this approach in the study of all the standard discrete distributions in Units 8 and 9. We shall study their properties in terms of their general parameter or parameters, without specifying their numerical values.

If X has the Bernoulli distribution given by (1), then

$$E(X) = 0.(1 - p) + 1.p = p$$

and

$$\text{Var}(X) = 0^2 \cdot (1 - p) + 1^2 \cdot p - p^2$$
$$= p(1 - p).$$

Notice that Var (X) ≤ E(X).

Its moment generating function (m.g.f.) is

$$M_x(t, p) = E[e^{tx}] = (1 - p) + pe^t,$$

which is valid for all real t and for $0 \leq p \leq 1$. We have deliberately introduced p in the symbol $M_x(t, p)$ for the m.g.f. of X to emphasise its dependence on the parameter p of the distribution.

The Bernoulli distribution is useful whenever the random experiment has only two possible outcomes, which may be labelled as success and failure. In the two situations discussed in the Introduction, we could identify success and failure. But in both these situations we were interested in a specified number of repetitions of the experiment. In other words, in each case, we were interested in the distribution of the sum of independent Bernoulli variates with the same value p of the parameter. We discuss this in the next section.

8.3 THE BINOMIAL DISTRIBUTION

In this section we are going to talk about the distribution of the sum of independent Bernoulli variables. You will see, in Theorem 1, that such a sum has a binomial distribution. But what is a binomial distribution?

We begin with the following definition.

Definition 1 : We say that a random variable X has a **binomial distribution** with parameters (n, p) if its p.m.f. is given by

$$b(j; n, p) = P[X = j] = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, 1, \ldots, n, \qquad \ldots (2)$$

where n is a positive integer and $0 < p < 1$.

Why is it called a **binomial** distribution? It's because b(j; n, p) is the (j + 1)th term in the binomial expansion of $\{p + (1 - p)\}^n$. This observation also leads to the conclusion that

$$\sum_{j=0}^{n} b(j; n, p) = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j}$$

$$= (1 - p + p)^n = 1,$$

which is as it should be.

We can interpret binomial distribution as the distribution of the total number of successes in n independent trials, each with the same probability p, of success. With this interpretation you will see that there are many situations in which this distribution can be applied. We have given some such situations in the examples a little later. Now, suppose $X_1, X_2, \ldots, X_n$ are independent Bernoulli r.vs. with the same p.m.f.,

$$P[X_j = 0] = 1 - p, P[X_j = 1] = p, j = 1, \ldots, n.$$

We have discussed independent trials in Sec. 6.6

We may identify a success at the j-th trial with the even $[X_j = 1]$ and a failure at the j-th trial with the event $[X_j = 0]$. Then $X = X_1 + \ldots + X_n$ is the total number of successes in n trials. To understand this, let us consider the coin-tossing experiment again. Suppose we toss an unbiased coin 5 times, i.e. n = 5. Now, the result of each toss could be either H or T. Suppose we call the result H a success. Then H at the jth toss is equivalent to $X_j = 1$ and T at the jth toss is equivalent to $X_j = 0$, j = 1, 2, . . . 5. So, if $X = X_1 + X_2 + \ldots X_5$, and if we get H in the first, second and the fifth toss and T in the rest, then X takes the value

x = 1 + 1 0 + 0 + 1 = 3, which is the number of Hs, or successes, in the 5 tosses.

Now let us obtain P[X = j] = 0, 1, . . ., n.

Notice that the sum $X_1 + \ldots + X_n$ equals j iff j or the X_i's are equal to 1 and the remaining (n – j) are all equal to zero. The probability that a specific set of j X_i s equal one and the remaining X_i s equal zero is $p^j(1 - p)^{n-j}$. This is so because there is one factor p for each

X_i ... and one factor $(1-p)$ for each X_i which is zero. The j factors p and $(n-j)$ factors ... p) get multiplied because of independence. However, there are $\binom{n}{j}$ mutually exclusive ways of choosing the j X_i s which equal one, the rest, $n-j$, of the X_is being equal to zero. Hence by the finite additivity property (P7 of Sec. 6.2.2),

$$P[X = j] = \binom{n}{j} p^j (1-p)^{n-j}, \; j = 0, 1, \ldots, n.$$

Hence the distribution of the total number of successes under the above conditions is the binomial distribution.

We have thus proved the following theorem.

Theorem 1: Let $X_1, \ldots X_n$ be n independent Bernoulli r.vs. with common p.m.f.

$$P[X_j = 1] = p, \; P[X_j = 0] = 1-p, \; j = 1, 2, \ldots, n.$$

... < 1. Then $X = X_1 + \ldots + X_n$ has binomial distribution with parameters n and p, ...ed by (2).

... this interpretation of a binomial distribution, let us look at some situations where this ... ation is useful.

Example 1 : A machine produces identical units. The proportion of defective units produced by the machine is known to be 1/20. We also know that successive units are statistically independent. Let us obtain the probability that in a sample of 10 units, there are at most 2 defectives.

If X denotes the number of defectives in a sample of 10 units, then X has binomial distribution with n = 10 and p = 1/20. Hence,

$$\begin{aligned} P[X \le 2] &= P[X=0] + P[X=1] + P[X=2] \\ &= b(0; 10, 1/20) + b(1; 10, 1/20) + b(2; 10, 1/20) \\ &= \binom{10}{0}(1/20)^0 (19/20)^{10} + \binom{10}{1}(1/20)(19/20)^9 \\ &\quad + \binom{10}{2}(1/20)^2 (19/20)^8 \\ &\approx 0.99. \end{aligned}$$

Example 2: The probability that a person recovers from a serious disease is 0.40. Let's find the probability that at least one of the 8 persons admitted to a hospital will survive.

For this, let us assume that the recovery or otherwise of the B patients is independent of each other. Thus, we want to know $P[X \ge 1]$, when X has binomial distribution with n = 8 and p = 0.40.

Observe that

$$\begin{aligned} P[X \ge 1] &= 1 - P[X = 0] \\ &= 1 - \binom{8}{0}(0.40)^0 (0.60)^8 \\ &= 1 - 0.017 \\ &= 0.983 \end{aligned}$$

Have you understood how we have solved these examples? See if you can solve some on your own now.

E1) Ten workers use electric power intermittently. Each worker has the same probability p = 1/5 of requiring a unit of power. If they work independently, find the probability that six or more workers require electric power simultaneously. If the supply is adjusted to five power units, this is the probability that the system would be overloaded.

E2) How many independent trials each with p = 0.01 must be performed to ensure that the probability of at least one success is 0.60 or more?

The calculation of probabilities associated with the binomial distribution is often complex. We now ask you to prove a result which is quite useful in this connection.

E3) Prove that

a) $b(j; n, p) = b(n - j; n, 1 - p)$,

b) $b(j + 1; n, p) = \frac{(n - j)p}{(j + 1)(1 - p)} b(j; n, p)$

We can use this result to calculate $b(j; n, p)$ recursively), starting with $b(0; n, p)$.

Now let us find the mean and variance of the binomial distribution.

Theorem 2: If X has binomial distribution with parameters n and p, then

$$E(X) = np, \quad Var(X) = np(1 - p).$$

Proof: By definition

$$E(X) = \sum_{j=0}^{n} jb(j; n, p)$$

$$= \sum_{j=1}^{n} j \frac{n!}{j!(n - j)!} p^j (1 - p)^{n-j},$$

where we have omitted the term corresponding to $j = 0$, since it is zero. Simplifying by using the relations $n! = n\{(n - 1)!\}$ and $\frac{j}{j!} = \frac{1}{(j - 1)!}$, we have

$$E(X) = np \sum_{j=1}^{n} \frac{(n - 1)!}{(j - 1)!\{n - 1 - (j - 1)\}!} p^{j-1} (1 - p)^{n-1-(j-1)}$$

$$= np \sum_{r=0}^{n-1} \binom{n-1}{r} p^r (1 - p)^{n-1-r}, \text{ where } r = j - 1.$$

$$= np(1 - p + p)^{n-1}$$

$$= np.$$

Now let's compute the variance.

You know that

$$Var(X) = E(X^2) - [E(X)]^2.$$

Since we have already computed E(X), we can find out Var(X) if we are able to calculate $E(X^2)$. The computation of $E(X^2)$ is simplified if we use the fact,

$$E(X^2) = E[X(X - 1)] + E(X).$$

Now,

$$E[X(X - 1)] = \sum_{j=0}^{n} j(j - 1) b(j; n, p)$$

$$= \sum_{j=2}^{n} j(j - 1) \frac{n!}{j!(n - j)!} p^j (1 - p)^{n-j},$$

since the first two terms corresponding to $j = 0$ and $j = 1$, vanish.

$$= n(n - 1) p^2 \sum_{j=2}^{n} \frac{(n - 2)!}{(j - 2)!\{n - 2 - (j - 2)\}!} p^{j-2} (1 - p)^{n-2-(j-2)}$$

$$= n(n-1)\,p^2 \sum_{r=0} \binom{n-2}{r} p^r (1-p)^{n-2-r}, \text{ where } r = j-2.$$

$$= n(n-1)p^2$$

Have you noticed that in this computation we have carried out simplifications which are similar to the ones used in the computation of E(X)? Finally

$$\text{Var}(X) = E[X(X-1)] + E(X) - \{E(X)\}^2$$

$$= n(n-1)p^2 + np - (np)^2$$

$$= np(1-p),$$

as required.

There is an easy consequence of this theorem, which we would like you to prove now.

E4) If $Y = X/n$ denotes the proportion of successes in n independent Bernoulli trials with constant probability p of success, then

$$E(Y) = p, \text{ Var}(Y) = \frac{p(1-p)}{n}.$$

E5) Use the results about the mean and variance of the sum of n independent r.vs. in Unit 7 for an alternative derivation of the mean and variance of a binomial r.v.

We conclude our discussion of the binomial distribution by obtaining its m.g.f.

Theorem 3: The moment generating function $M_X(t)$ of the binomial distribution with parameters n and p is

$$M_X(t) = \{1 + p(e^t - 1)\}^n.$$

Proof: By definition

$$M_X(t) = E[\exp(tX)]$$

$$= \sum_{j=0}^{n} e^{jt} \binom{n}{j} p^j (1-p)^{n-j}$$

$$= \sum_{j=0}^{n} \binom{n}{j} (pe^t)^j (1-p)^{n-j}$$

$$= \{pe^t + 1 - p\}^n$$

$$= \{1 + p(e^t - 1)\}^n,$$

which is the required result.

From the m.g.f. also you can see that $E(X) = np$ and $\text{Var}(X) = np(1-p)$.

We now prove that the sum of two independent binomial variates with common probability 'p' of success is again a binomial variate.

Corollary: Let X and Y be independent binomial variates with parameters (n, p) and (m, p), respectively. Then X+Y has a binomial distribution with parameters (m + n, p).

Proof: Now, X and Y can be regarded as the sum of n and m independent Bernoulli variates, respectively. Suppose

$$X = X_1 + X_2 + \ldots + X_n \text{ and } Y = X_{n+1} + X_{n+2} + \ldots + X_{n+m}$$

Then $X + Y = X_1 + X_2 + \ldots + X_{n+m}$ and X_i, $i = 1, 2, \ldots, n+m$ are independent.

Thus, X + Y is a sum of n+m independent Bernoulli variates with probability, p, of success. Hence, X + Y has binomial distribution with parameters (n + m, p).

We have mentioned earlier that a binomial distribution is the distribution of the total number of successes in n trials, each with the same probability of success, where

1) each trial can result in two mutually exclusive outcomes and

2) successive trials are independent.

Now, there are two ways in which we can generalise the binomial distribution. We can assume that either

1) each trial can result in more than two mutually exclusive outcomes, or

2) successive trials are not independent.

We shall follow the first approach in the next section and the second approach in Sec. 8.5

8.4 THE MULTINOMIAL DISTRIBUTION

Sometimes we come across situations where a trial of an experiment may result in more than two outcomes. Here are some examples of such situations.

i) A group of 100 persons is classified according to their blood- groups O, A, B and AB. Let r_1, r_2, r_3 and r_4 denote the number of persons with the blood groups O, A, B and AB, respectively. Then r_1, r_2, r_3 and r_4 are non-negative integers with $r_1 + r_2 + r_3 + r_4 = 100$. Here each person can be classified into one and only one of the $k = 4$ classes.

ii) In a game of bridge, the 52 playing cards are divided amongst $k = 4$ players such that each player gets 13 cards.

iii) The population of a town can be classified into $k = 21$ different age groups, 0 – 2, 3 – 7, 8 – 12, . . ., 98 and above.

iv) The teachers in a university can be classified into $k = 3$ categories; lecturer, reader and professor.

v) In a Lok Sabha constituency, there are 5 candidates. Before the polling date, the voters can be classified into six classes, five according to their choice of the candidate, the sixth class being of those who are still undecided.

To deal with such situations, we first need to find the total number of ways in which n distinct objects can be classified into k different classes so that r_1 belong to Class 1, r_2 belong to Class 2, . . . , and r_k to Class k. Of course, it is necessary to have $r_1 + r_2 + \ldots + r_k = n$.

You are already familiar with the case, $k = 2$.

When $k = 2$, we can classify, n objects into two classes such that r_1 belong to Class 1 and $r_2 = (n - r_1)$ belong to Class 2 in

$$\binom{n}{r_1} = \frac{n!}{r_1!\,(n - r_1)!} = \frac{n!}{r_1!\,r_2!}$$

ways. This is so because we can choose r_1 objects out of n objects in $\binom{n}{r_1}$ ways and every such choice leaves a unique group of $n - r_1 = r_2$ objects which belong to Class 2.

We now generalise this argument in the following theorem.

Theorem 4 : The number of ways of classifying n distinct objects in k classes, such that r_1 belong to Class 1, r_2 belong to Class 2, . . ., r_k belong to Class k, subject to the condition $r_1 + r_2 + \ldots + r_k = n$, is

$$\frac{n!}{r_1!\,r_2!\ldots r_k!}$$

Proof : We know that the r_1 objects belonging to Class 1 can be chosen in $\binom{n}{r_1}$ ways out of the n objects. Having chosen these r_1 objects, the r_2 objects that are to be assigned to Class 2

can be selected out of the remaining $(n - r_1)$ objects in $\binom{n-r_1}{r_2}$ ways. The number of ways of selecting r_3 objects for Class 3 out of the remaining $(n - r_1 - r_2)$ objects is $\binom{n-r_1-r_2}{r_3}$. We continue this procedure. So, having put objects in Classes $1, 2, \ldots, j-1$, the r_j objects in Class j can be selected out of the balance of $n - r_1 - r_2 - \ldots - r_{j-1}$ objects in

$$\binom{n - r_1 - r_2 - \ldots - r_{j-1}}{r_j}$$

ways. Hence the required number is given by

$$\binom{n}{r_1}\binom{n-r_1}{r_2}\cdots\binom{n-r_1-r_2-\ldots-r_{k-1}}{r_k}$$

$$= \frac{n!}{r_1!\,(n-r_1)!} \times \frac{(n-r_1)!}{r_2!\,(n-r_1-r_2)!} \times \ldots \times \frac{(n-r_1-\ldots-r_{k-1})!}{r_k!\,0!}$$

$$= \frac{n!}{r_1!\,r_2!\ldots r_k!}$$

The proof is complete.

Now consider n independent trials, each of which results in one of the k possible outcomes. Suppose the probabilities of these outcomes are $p_1, p_2, \ldots, p_k$, respectively. Then $\sum_{j=1}^{k} p_j = 1$. Let $X_1, X_2, \ldots, X_k$ be the respective frequencies of k outcomes. Then $X_1, X_2, \ldots, X_k$ assume non-negative integral values. Now, we wish to find the probability

Compare this with the discussion just before Theorem 1.

$$P[X_1 = r_1, X_2 = r_2, \ldots, X_k = r_k],$$

where $\{r_1, r_2, \ldots, r_k\}$ is a fixed set of non-negative integers adding up to n.

Consider a specific sequence of outcomes resulting in r_j outcomes of the jth type, $j = 1, 2, \ldots, k$. The probability that r_1 is the frequency of the first outcome, r_2, that of the second outcome, $\ldots, r_k$, that of the k^{th} outcome, is $p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k}$.

But, according to Theorem 4, there are $\frac{n!}{r_1!\,r_2!\ldots r_k!}$ such sequences, where we have r_1 as the frequency of the first outcome, r_2 as the frequency of the second outcome, $\ldots, r_k$ as that of the k^{th} outcome. Therefore,

$$P[X_1 = r_1, X_2 = r_2, \ldots, X_k = r_k] = \frac{n!}{r_1!\,r_2\ldots r_k!}\, p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k}$$

for all non-negative integral $r_1, \ldots, r_k$ such that $r_1 + \ldots + r_k = n$.

This leads to the following definition.

Definition 2: The r.vs. $X_1, \ldots X_k$ are said to have a **multinomial distribution** with parameters $(n; p_1, p_2, \ldots, p_k)$, if their joint p.m.f. is

$$f(r_1, r_2, \ldots, r_k; n, p_1, \ldots, p_k)$$

$$= P[X_1 = r_1, \ldots, X_k = r_k]$$

$$= \frac{n!}{r_1!\,r_2!\ldots r_k!}\, p_1^{r_1} p_2^{r_2} \ldots p_k^{r_k}$$

for $r_i = 0, 1, \ldots, n$ subject to $\sum_{i=1}^{k} r_i = n$, $p_j \geq 0$, and $\sum_{i=1}^{k} p_i = 1$.

This distribution is called 'multinomial', because the terms of the p.m.f. are the corresponding terms of the multinomial expansion of $(p_1 + \ldots + p_k)^n$.

We now give two examples where the variables have a multinomial distribution. Let us use the p.m.f. of a multinomial distribution to find the probabilities in these examples.

Example 3 : In a population, 43% have blood group 0, 45% have A, 8% have B and 4% have blood group AB. Sixteen persons belonging to this population are classified according to their blood group. Let us find the probability that there will be 4 of each type.

Here $k = 4$, $p_1 = 0.43$, $p_2 = 0.45$, $p_3 = 0.08$, $p_4 = 0.04$ and $n = 16$. We want to find

$$P[X_1 = 4, X_2 = 4, X_3 = 4, X_4 = 4]$$

$$= \frac{16!}{4!\,4!\,4!\,4!} (0.43)^4 (0.45)^4 (0.08)^4 (0.04)^4$$

$$\approx 9 \times 10^{-6}.$$

Example 4: The probabilities are 0.40, 0.50 and 0.10, that an electric bulb will last for at most 220 days, 221 to 260 days and 261 days or more, respectively. Suppose 10 bulbs are picked up one by one from a manufacturing line. Suppose we want to find the probability that among 10 such bulbs, three will last for 220 days or less, six will serve for 221 to 260 days and one will last for more than 261 days.

Let us assume that these selections constitute independent multinomial trials.

Here $k = 3$, $n = 10$, $p_1 = 0.40$, $p_2 = 0.50$ and $p_3 = 0.10$. We have to find

$$P[X_1 = 3, X_2 = 6, X_3 = 1]$$

$$= \frac{10!}{3!\,6!\,1!} (0.40)^3 (0.50)^6 (0.10)^1$$

$$= 0.084.$$

The study of the multinomial distribution is a little more complicated than that of the binomial distribution since the multinomial distribution specifies the joint distribution of k r.vs. We give only the means, variances and covariances for the multinomial r.vs. in the following theorem without proving it. We are sure you will be able to prove it.

Theorem 5: Let the joint distribution of $X_1, \ldots, X_k$ be a multinomial distribution with parameters $(n; p_1, \ldots, p_k)$. Then

$$E[X_j] = np_j, \text{ Var } (X_j) = np_j (1 - p_j)$$

$$\text{Cov } (X_i, X_j) = -np_ip_j, \; i \neq j,$$

$$i, j = 1, \ldots, k.$$

Here are a few exercises which you should solve.

E6) Twelve unbiased dice are rolled. What is the probability that each of the six faces occurs twice?

Hint : $k = 6$, $n = 12$, $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$

E7) An item is defective with probability p and good with probability $q = 1 - p$. It may be inspected with probability p' and may not be inspected with probability $q' = 1 - p'$. Assume that these probabilities are the same for all the items, which are independent. Assume further that the decision to inspect or not to inspect an item is made without the knowledge of its quality. Below we give the k categories and the corresponding probabilities that an item belongs to them.

Category	Probability
Good item is inspected	qp'
Good item is not inspected	qq'
Defective item is inspected	pp'
Defective item is not inspected	pq'

Find the probability that of 10 such items all are defective and are inspected.

E8) Show that the marginal distribution of X_j is binomial with parameters (n, p_j).

Hint : Use the argument preceding Definition 2.

E9) Prove Theorem 5.

(**Hint :** Note that

i) For a fixed i, X_i is a binomial r.v. with parameters n and p_i.

ii) For a fixed pair (i, j), $X_i + X_j$ also is a binomial r.v.

Now use Theorem 2 and the result about Var(X+Y) from Unit 7.)

In this section we have generalised the binomial distribution by assuming that each trial results in more than two outcomes. In the next one we shall see what happens if we assume that successive trials are not independent. This leads us to the hypergeometric distribution.

8.5 THE HYPERGEOMETRIC DISTRIBUTION

The hypergeometric distribution deals with trials, each of which has exactly two possible outcomes. But so does the binomial distribution. Then what is the difference between them? Let us see.

Suppose we have 20 items, of which 4 are defective and the remaining 16 are good items. Then, if we select an item at random, the probability that it is defective is 4/20 = 1/5, and that it is good is 16/20 = 4/5. Now, if we replace the selected item after every selection, these probabilities would remain unchanged at successive selections. We may also assume that successive selections are independent. Suppose we are interested in the number X of defective items in 10 such independent selections made with replacement. Then X has a binomial distribution with n = 10, p = 1/5.

Now, can you find the probability that in 10 such selections, there are 2 defective and 8 good items? It is given by

$$b(2; 10, 1/5) = \binom{10}{2}(1/5)^2 (4/5)^8 \approx 0.302.$$

Now suppose we make a slight change in our trials. We, now, select the items without replacement. What difference does this make? The probability that the first item is defective continues to be 1/5. But the probability that the second item is defective, given that the first item was defective, is 3/19, which is different from 1/5. It also implies that the successive selections are no longer independent.

How do we find the probability that in 10 selections made at random without replacement, there are 2 defective and 8 good items? Let's see. Note that the total number of ways of selecting 10 items out of 20 items is $\binom{20}{10}$. The number of ways of selecting 2 defectives out of 4 defectives is $\binom{4}{2}$ and that of selecting 8 good items out of 16 good items is $\binom{16}{8}$. Hence, the total number of ways of selecting 2 defective and 8 good items out of 4 defective and 16 good ones is $\binom{4}{2}\cdot\binom{16}{8}$. Further, since all the selection are at random, they are equally likely. Hence, the required probability is

$$\frac{\binom{4}{2}\binom{16}{8}}{\binom{20}{10}} = 0.418.$$

In this case, we find that the probability of getting 2 defective and 8 good items is

- 0.302, if the trials are independent, and
- 0.418, if the trials are not independent.

We now generalise the above argument and consider a set of N items of which M are defective and N – M are good. We select n items **without** replacement and want to find the probability that j of them are defective. Observe that when j of n items are defective, the remaining n – j must be good items. But since there are M defective and N – M good items, we must also have $j \leq M$, $n - j \leq N - M$

We can choose n items out of N in $\binom{N}{n}$ ways. The j defectives can be chosen out of M defectives in $\binom{M}{j}$ ways and the n – j good items can be chosen out of N – M good ones in $\binom{N-M}{n-j}$ ways. Hence, X denotes the number of defective items selected,

$$P[X = j] = \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

for $j = 0, 1, \ldots, n, j \leq M, n - j \leq N - M$.

This discussion leads us to the following definition.

Definition 3: A r.v. X has the **hypergeometric distribution** with parameters (n, N, M) if its p.m.f. is

$$h(j; n, N, M) = \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

for $j = 0, 1, \ldots, n, j \leq M, n - j \leq N - M$. Here n, N, M are positive integers, $n \leq N, M \leq N$.

Now, we give examples of two situations where the variables have hypergeometric distribution.

Example 5 : A quality control engineer inspects two randomly selected units from a lot of 20 units. If both the units are in working condition, the lot is accepted. Otherwise all the remaining units are inspected. Let us find the probability that a lot of 20 units containing 8 defective units is accepted without further inspection.

Here N = 20, M = 8, and n = 2. We seek the probability

$$P[X = 0] = h(0; 2, 20, 8)$$

$$= \frac{\binom{8}{0}\binom{12}{2}}{\binom{20}{2}} = 0.347.$$

Example 6 : A lake contains N fish. Suppose we catch 1000 of these fish, mark each of them with a red spot and release them in the lake. The next day, we make a new catch of 1000 fish and find that 100 of these have red spots. Let us find the probability of this occurrence.

The probability that we wish to find is

$$P[X = 100] = h(100; 1000, N, 1000)$$

$$= \frac{\binom{1000}{100}\binom{N-1000}{900}}{\binom{N}{1000}}$$

Of course, we cannot numerically evaluate this without a knowledge of N. However, the above discussion is useful in developing methods for estimation of the sizes of mobile biological populations. The method described above is called, for obvious reason, the capture-recapture method.

Let us assume that we know the value of N, say N = 2000. Then

$$P[X = 100] = \frac{\binom{1000}{100}\binom{1000}{900}}{\binom{2000}{1000}}$$

You will agree that the exact computation of this probability is very cumbersome.

This is one of the main difficulties with the use of the hypergeometric distribution. The computation of its probabilities can be quite tedious, especially if N and M are large. However, suppose n is small compared to N, (n/N < 0.05, say) then there is not much difference between sampling with and without replacement. In fact, we can replace h(j; n, N, M) by the binomial probability b(j; n, p) where $p \cong M/N$.

We conclude this section with the evaluation of the mean and variance of the hypergeometric distribution.

Theorem 6 : If X has the hypergeometric distribution with parameters n, N and M, then

$$E(X) = \frac{nM}{N} \text{ and } Var(X) = \frac{nM(N-M)(N-n)}{N^2(N-1)}.$$

Proof: We evaluate the mean E(X) directly as follows:

$$E(X) = \sum_{j=0}^{n} j\, h(j; n, N, M)$$

$$= \sum_{j=1}^{n} j \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

$$= \sum_{j=1}^{n} \frac{M!}{(j-1)!\,(M-j)!} \frac{\binom{N-M}{n-j}}{\binom{N}{n}}$$

$$= \frac{M}{\binom{N}{n}} \sum_{j=1}^{n} \frac{(M-1)!}{(j-1)!\,\{(M-1)-(j-1)\}!} \binom{N-M}{n-1-(j-1)}$$

$$= \frac{M}{\binom{N}{n}} \sum_{r=0}^{n-1} \binom{M-1}{r}\binom{N-1-(M-1)}{n-1-r}$$

Now,

$$\sum_{j=0}^{r} \binom{m}{j}\binom{n}{r-j} = \binom{m+n}{r}.$$

Expand both sides of $(1+x)^{m+n} = (1+x)^m \cdot (1+x)^n$, and compare the coefficients of x^r to get this result.

Using this result we get

$$E(X) = \frac{M\binom{N-1}{n-1}}{\binom{N}{n}}$$

$$= \frac{nM}{N}.$$

In order to compute variance, we adopt the method that was used in the evaluation of the variance of the binomial distribution (see Theorem 2). That is, we use the relation,

$$Var(X) = E[X(X-1)] + E(X) - \{E(X)\}^2.$$

Since we already know E(X), the only thing that remains to be done is to find E[X(X – 1)]. We are sure you can handle that. So we are leaving it to you as an exercise.

When N and M are large, check that $E(X) \cong np$ and $Var(X) \cong np(1-p)$, which are the mean and variance of a binomial r.v.

E10) Find Var(X), where X has hypergeometric distribution.

E11) A parcel of 20 books contains 5 books with loose bindings. What is the probability that a random selection of 10 of the 20 books, drawn without replacement, will contain the five books with loose binding?

E12) Show that

$$h(j+1; n, N, M) = \frac{(n-j)(M-j)}{(j+1)(N-M-n+j+1)} h(j; n, N, M).$$

(This is a recurrence relation which can be used to compute the hypergeometric probabilities)

That brings us to the end of this unit. In the next unit we'll take up the study of some more, frequently used, discrete probability distributions. But before that, let us briefly recall what we have studied in this unit.

8.6 SUMMARY

In this unit we have covered the following points.

1) If the r.v. X assumes only two values, 0 and 1, with $P[X = 0] = 1 - p$ and $P[X = 1] = p$, $0 \le p \le 1$, the X has a **Bernoulli distribution.**

2) The **binomial distribution** gives the probability distribution of the total number of successes in n independent Bernoulli trials with constant chance p of success in each trial. It can also be regarded as the distribution of the sum of n independent and identically distributed Bernoulli r.vs.

   If X has a binomial distribution with parameters (n, p), then $E(X) = np$, $Var(X) = np(1-p)$.

3) The natural extension of the binomial distribution is the **multinomial distribution** when each trial results in $k > 2$ disjoint outcomes. The binomial distribution is a particular case of the multinomial distribution with $k = 2$. If $X_1, X_2, \ldots X_k$ are r.vs. with a multinomial distribution with parameters $(n; p_1, p_2, \ldots p_k)$, then $E(X_j) = np_j$, $Var(X_j) = np_j(-p_j)$, $Cov(X_i, X_j) = -np_ip_j$, $i \neq j$, $i, j = 1, 2, \ldots, k$.

4) We obtain the **hypergeometric distribution** when we select n objects without replacement out of N objects, of which M are of type 1 and N – M are of type 2. The distribution of the number of type 1 objects so selected is the hypergeometric distribution. If $n/N < 0.05$, the binomial probabilities $b(j; n, M/N)$ closely approximate the hypergeometric probabilities $h(j; n, N, M)$.

$$E(X) = \frac{nM}{N}, \quad Var(X) = \frac{nM(N-M)(N-n)}{N^2(N-1)}$$

5) The binomial distribution has the so called reproductive property, viz., if X and Y are independent binomial variates with parameters (n, p) and (m, p), respectivly, then X + Y has binomial distribution with parameters (m + n, p).

8.7 SOLUTIONS AND ANSWERS

E1) The number X of workers using electricity simultaneously has binomial distribution with n=10, p=1/5. We need to find

$$P[X \ge 6] = \sum_{j=6}^{10} b(j; 10, 1/5)$$

$$\cong 0.0064$$

E2) The probability of at least one success in n independent Bernoulli trials is

$$1 - b(0, n, p) = 1 - (1-p)^n.$$

We have $p = 0.01$ and we want $1 - (0.99)^n \ge 0.60$ or $n \ge 92$.

E3) a) $$b(j; n, p) = \binom{n}{j} p^j (1-p)^{n-j}$$

$$= \binom{n}{n-j} (1-p)^{n-j} \{1 - (1-p)\}^{n-(n-j)}$$

$= b(n-j; n, 1-p).$

b) $b(j+1; n, p) = \binom{n}{j+1} p^{j+1} (1-p)^{n-j-1}$

$= \frac{n!}{j!\,(n-j)!} \cdot \frac{(n-j)}{(j+1)} p^j (1-p)^{n-j} \cdot \frac{p}{1-p}$

$= \frac{(n-j)p}{(j+1)(1-p)} \binom{n}{j} p^j (1-p)^{n-j}$

$= \frac{(n-j)p}{(j+1)(1-p)} b(j; n, p).$

E4) From Unit 7 we know that

$E(aX + b) = aE(X) + b$ and $\text{Var}(aX + b) = a^2 \text{Var}(X)$

$\therefore$ The result follows.

E5) If X is a binomial r.v. with parameters (n, p) then X can be considered as the sum of n independent Bernoulli r.vs, $X_1, X_2, \ldots, X_n$, with common parameter p.

Therefore

$$E(X) = E(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} E(X_i) = np \text{ and}$$

$$\text{Var}(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i)$$

$$= np(1-p).$$

E6) We need to compute

$$\frac{12!}{(2!)^6} (1/6)^{12} = 0.0034.$$

E7) The required probability is

$$\frac{10!}{10!\,0!\,0!\,0!} (qp')^0 (qq')^0 (pp')^{10} (pq')^0 = (pp')^{10}.$$

E8) If we treat classification in Class j as "success" and classification into any other class as "failure", the probability of success is p_j and we have n independent Bernoulli trials each with probability p_j of success. The result follows.

E9) For fixed j, X_j is a binomial r.v. with parameters (n, p_j)

$E[X_j] = np_j$, $\text{Var}(X_j) = np_j(1-p_j)$.

For $i \neq j$, $\text{Var}(X_i + X_j) = \text{Var}(X_i) + \text{Var}(X_j) + 2\,\text{Cov}(X_i, X_j)$

$n(p_i + p_j)(1 - p_i - p_j) = np_i(1-p_i) + np_j(1-p_j) + 2\text{Cov}(X_i, X_j)$

$$\text{Cov}(X_i, X_j) = \frac{n}{2}\left[(p_i + p_j)(1 - p_i - p_j) - p_i(1-p_i) - p_j(1-p_j)\right]$$

$$= -np_ip_j.$$

E10) $E[X(X-1)] = \sum_{j=0}^{n} j(j-1)\, h(j; n, N, M)$

$$= \sum_{j=2}^{n} j(j-1) \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$$

$$= \sum_{j=2}^{n} \frac{M!}{(j-2)!\,(M-j)!} \frac{\binom{N-M}{n-j}}{\binom{N}{n}}$$

$$= \frac{M(M-1)}{\binom{N}{n}} \sum_{j=2}^{n} \frac{(M-2)!}{(j-2)!\,(M-j)!} \binom{N-M}{n-2-(j-2)}$$

$$= \frac{M(M-1)}{\binom{N}{n}} \sum_{r=0}^{n-2} \binom{M-2}{r}\binom{N-M}{n-2-r}$$

$$= \frac{M(M-1)}{\binom{N}{n}} \cdot \binom{N-2}{n-2}$$

$$= \frac{M(M-1)\ n(n-1)}{n\,(N-1)}$$

E11) We need to compute

$$h(5; 10, 20, 5) = \frac{\binom{15}{5}\binom{5}{5}}{\binom{20}{50}}$$

$$\cong 0.016.$$

E12) We have

$$h(j+1; n, N, M) = \frac{\binom{M}{j+1}\binom{N-M}{n-j-1}}{\binom{N}{n}}$$

$$= \frac{M!}{(j+1)!\,(M-j-1)!} \frac{(N-M)!}{(n-j-1)!\,(N-M-n+j+1)!} \Big/ \binom{N}{n}$$

$$= \frac{(M-j)}{(j+1)} \frac{M!}{j!\,(M-j)!} \frac{(N-M)!}{(n-j)!\,(N-M-n+j)!} \frac{n-j}{(N-M-n+j+1)} \Big/ \binom{N}{n}$$

$$= \frac{(M-j)\,(n-j)}{(j+1)\,(N-M-n+j+1)} \binom{M}{j} \binom{N-M}{n-j} \Big/ \binom{N}{n}$$

which is the required result.

UNIT 9 STANDARD PROBABILITY DISTRIBUTIONS : PART - II

Structure

9.1 INTRODUCTION

The standard probability distributions that we studied in Unit 8 are all distributions of r.vs. which assume a finite number of values. However, there are many situations of practical as well as theoretical interest which require the use of r.vs. whose values can be arranged in an unending sequence. The simplest such cases are of those r.vs. which assume the values 0, 1, 2, . . . , i.e., those which are non-negative, integer-valued r.vs.

The usual coin tossing experiment provides an example of this type. Suppose we toss a coin until a head turns up, and denote by X the number of tosses required for the purpose. Then $X = 1, 2, \ldots$, and, in general, we cannot specify an upper bound k such that $P[X \leq k] = 1$.

An obvious extension of the above example is the following. Suppose we decide to toss the coin until a specified number, r say, of heads turn up. In this situation, the number X of tosses required is $r, r + 1, r + 2, \ldots$.

Although both these illustrations seem mainly to be of theoretical interest, they are useful in many statistical and probabilistic problems of an advanced nature. Since they are concerned with waiting times (number of trials) required for the first or r-th occurrence of a specific event, the associated distributions are called **waiting time distributions.** We shall discuss two simple waiting time distributions in this unit : the geometric distribution and the negative binomial distribution.

The situation described below is of a different type. Nevertheless, it also leads to a r.v. with infinitely many values.

A radioactive substance emits particles called α- particles. The number of α-particles emitted during a time interval of one hour, say, can be recorded by an instrument. The number X of such particles can be $0, 1, 2, \ldots$. The r.v. in this follows Poisson distribution. In this unit we shall also be discussing the properties of the Poisson distribution.

Objectives

After reading this unit you should be able to :

- define the geometric, negative binomial and Poisson distributions
- calculate the mean and variance of these distributions
- compute probabilities of events associated with these standard distributions.

9.2 THE GEOMETRIC DISTRIBUTION

In this section we'll discuss the geometric distribution. Let us see first how suh a distribution arises.

Let p denote the probability of a success in a Bernoulli trial, $0 < p < 1$. Consider independent repetitions of such a trial. Denote by X, the number of trials required for first success. Then the r.v. X takes the values 1, 2, 3, . . . and by definition

$$P[X = 1] = p.$$

In order to obtain $P[X = j]$ for $j \geq 2$, observe that the event $[X = j]$ occurs iff, the first $j - 1$ trials result in a failure and the j-th trial is a success. The probability that we have first $(j - 1)$ failure followed by a success is

$$P[X = j] = (1 - p)^{j-1}p,$$

by virtue of independence of the repeated Bernoulli trials. The r.v X here is said to have a geometric distribution. Here is the formal definition.

Definition 1 : A r.v. X is said to have the geometric distribution with parameter p, $0 < p < 1$, if its p.m.f. is given by

$$P[X = j] = p(1 - p)^{j-1}, j = 1, 2, \ldots . \qquad \ldots (1)$$

The distribution derives its name from the fact that $P[X = j]$ is the j-th term of the geometric series

$$\sum_{j=1}^{\infty} p(1 - p)^{j-1}.$$

For $p \in]0, 1[$, the above infinite series is convergent and its sum is

$$\frac{p}{1 - (1 - p)} = 1.$$

which is what is required.

A slightly different way of arriving at the geometric distribution is to consider a sequence $\{Y_n, n \geq 1\}$ of independent and identically distributed Bernoulli r.vs., such that

$$P[Y_n = 1] = p, P[Y_n = 0] = 1 - p$$

for all $n \geq 1$. Identify $Y_n = 1$ with success at the n-th Bernoulli trial and $Y_n = 0$ with failure at the n-th trial. Then the event $[X = j]$ is the same as the event $[Y_1 = 0, \ldots, Y_{j-1} = 0, Y_j = 1]$, and hence, by virtue of independence of Y_n s,

Recall that if $a + ar + ar^2 + \ldots$ is a convergent geometric series, then its sum is $\frac{a}{1 - r}$.

$$P[X = j] = P[Y_1 = 0]\, P[Y_2 = 0] \ldots P[Y_{j-1} = 0]\, P[Y_j = 1]$$
$$= (1 - p)^{j-1}p, j = 1, 2, \ldots$$

Now let's see some examples of this distribution.

Example 1 : The probability is 0.70 that a candidate will pass an examination. Suppose we want to find the probability that he will pass the examination at the fourth attempt.

Assuming that the successive attempts of the candidate are independent repetitions of a Bernoulli trial with $p = 0.70$, the required probability is

$$P[X = 4] = 0.70\,(1 - 0.70)^3$$
$$= 0.0189.$$

Actually, the assumptions made in Example 1 are not very realistic. In particular, they imply that the candidate learns nothing from his first three failures.

Example 2 : Let $\{Y_n, n = 1, 2, \ldots\}$ be a sequence of independent and identically distributed r.vs. (i.i.d.r.vs.), such that for all $n \geq 1$,

$$P[Y_n = 0] = 1/4, P[Y_n = 1] = 1/4, P[Y_n = 2] = 1/2.$$

So, each Y_n can take the values 0, 1, 2. Consider a sequence of observed values of $Y_1, Y_2, \ldots$ Let X be the number of Y_n s that need to be observed to obtain the first 0 in this sequence. Let us find the probability that $X > 4$.

We say that a success occurs at trial number n if $Y_n = 0$. In view of the identical nature of the distribution of Y_n, the probability of a success at any trial is $p = 1/4$. The r.v. X therefore has the geometric distribution with $p = 1/4$. We need to compute

$$P[X > 4] = \sum_{j=5} P[X=j]$$

$$= \sum_{j=5}^{\infty} \frac{1}{4} (3/4)^{j-1}$$

$$= \frac{1}{4} \frac{(3/4)^4}{(1 - 3/4)}$$

$$= (3/4)^4$$

$$\cong 0.316.$$

Now here are some simple exercises for you to solve.

E1) Obtain the probability that in independent tosses of a balanced die, we will have to wait for at least 5 tosses to obtain the first six.

E2) Cards are drawn at random and with replacement from a well- shuffled pack of 52 playing cards. Find the probability that the first ace will appear before the fifth selection.

We shall now study the properties of the probability distribution of X specified by (1).

The following theorem gives the mean and variance of X.

Theorem 1 : If the r.v. X has geometric distribution with p.m.f. specified by (1), its mean and variance are

$$E(X) = \frac{1}{p}, \quad Var(X) = \frac{(1-p)}{p^2}$$

Proof : By definition

$$E(X) = \sum_{j=1}^{\infty} j\, p(1-p)^{j-1}$$

$$= p \sum_{j=1}^{\infty} j\, q^{j-1}, \text{ where } q = (1-p).$$

To sum the infinite series $S_1 = \sum_{j=1}^{\infty} j\, q^{j-1}$, note that

$$(1-q)S_1 = (1-q)(1 + 2q + 3q^2 + \ldots)$$

$$= 1 + q + q^2 + \ldots$$

$$= \frac{1}{1-q}$$

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

Hence. $S_1 = 1/p^2$ and therefore, $E(X) = 1/p$.

The variance, Var (X), will be obtained by employing the familiar technique of writing

$$Var(X) = E[X(X-1)] + E(X) - \{E(X)\}^2.$$

It is, therefore, enough to compute

$$E[X(X-1)] = \sum_{j=1}^{\infty} j(j-1)p\, q^{j-1}, \quad q = 1 - p$$

$$= p \sum_{j=2}^{\infty} j(j-1)\, q^{j-1}$$

In order to sum the infinite series

$$S_2 = \sum_{j=2}^{\infty} j(j-1)q^{j-1},$$

observe that

$$S_2 = \sum_{r=1}^{\infty} r(r+1)q^r, \text{ where } r = j-1.$$

Hence

$$(1-q)\,S_2 = \sum_{r=1}^{\infty} r(r+1)q^r - \sum_{r=1}^{\infty} r(r+1)\,q^{r+1}$$

$$= 2\sum_{r=1}^{\infty} rq^r$$

$$= 2q\sum_{r=1}^{\infty} rq^{r-1}$$

$$= 2qS_1$$

$$= \frac{2q}{p^2}, \text{ since we have already seen that } S_1 = \frac{1}{p^2}.$$

Therefore, $S_2 = \frac{2q}{p^3}$.

Thus, finally,

$$E[X(X-1)] = p\left(\frac{2q}{p^3}\right) = \frac{2q}{p^2},$$

and therefore,

$$\text{Var}(X) = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2}$$

$$= \frac{(1-p)}{p^2}$$

which completes the proof of the theorem.

We now obtain the moment generating function of the geometric distribution. We'll use it in the next section while discussing the so-called negative binomial distribution.

Theorem 2 : Let X be a r.v. with geometric distribution specified by the p.m.f. (1). Its m.g.f. is

$$M_x(t) = \frac{pe^t}{\{1-(1-p)e^t\}},$$

valid for all t such that $t < \ln\left(\frac{1}{1-p}\right)$.

Proof : By definition

$$M_x(t) = E[e^{tX}]$$

$$= \sum_{j=1}^{\infty} e^{tj}\,p(1-p)^{j-1}$$

$$= pe^t \sum_{j=1}^{\infty} \{e^t(1-p)\}^{j-1}$$

$$= pe^t \sum_{r=0}^{\infty} \{(1-p)e^t\}^r, \text{ where } r = j-1,$$

$$= \frac{pe^t}{\{1-(1-p)e^t\}},$$

which is valid only if $(1-p)e^t < 1$ or $t < \ln\left(\frac{1}{1-p}\right)$. This is so, because only when $t < \ln\left(\frac{1}{1-p}\right)$, the infinite series defining $M_x(t)$ is absolutely convergent.

We conclude this section with an interesting property of the geometric distribution.

Let X be a geometric r.v. with parameter p. Then for any positive integer j,

$$P[X > j] = \sum_{r=j+1}^{\infty} P[X = r]$$

$$= \sum_{r=j+1}^{\infty} p(1-p)^{r-1}$$

$$= p(1-p)^j \sum_{t=0}^{\infty} (1-p)^t, \; t = r-(j+1)$$

$$= (1-p)^j.$$

Consider the event $[X > j+k]$, where k is also a positive integer. Since $X > j+k$ implies that $X > j$,

$$[X > j+k] \cap [X > j] = [X > j+k].$$

Let us now evaluate the conditional probability that the waiting time for first success exceeds $j+k$, given that it exceeds j; i.e. we wish to evaluate $P[X > j+k \mid X > j]$. By definition,

$$P[X > j+k \mid X > j] = \frac{P[X > j+k, X > j]}{P[X > j]}$$

$$= \frac{P[X > j+k] \cap [X > j]}{P[X > j]}$$

$$= \frac{P[X > j+k]}{P[X > j]}$$

$$\frac{(1-p)^{j+k}}{(1-p)^j}$$

$$= (1-p)^k$$

$$= P[X > k].$$

Thus, we have shown that for all positive integers j and k

$$P[X > j+k \mid X > j] = P[X > k]$$

i.e., the conditional probability that the waiting time to first success exceeds $j+k$, given that it exceeds j, is the same as the probability that it exceeds k. In other words, the fact that we have waited for at least j trials for the first success does not affect the probability that we will have to wait for a further k trials. This property is therefore called the **lack of memory property** of the geometric distribution, or its **forgetfulness property.** In fact, the geometric distribution is the only distribution on the set of non-negative integers with the lack of memory property. This has important consequences in the study of more complicated systems of r.vs. called Markov chains. But we cannot go into its details in this course.

In this section we have seen how geometric distribution arises. We have also derived some properties of this distribution. In particular, we have noted that this is the only distribution with the forgetfulness property.

We'll take up the study of the negative binomial distribution in the next section.

9.3 THE NEGATIVE BINOMIAL DISTRIBUTION

This section discusses the properties of the so-called negative binomial distribution which is a generalisation of the geometric distribution. You know that the geometric distribution gives the distribution of the number of trials required to obtain the **first** success in independent repetitions of a Bernoulli trial. Now suppose we want to find the distribution of the number of trials required to obtain the **r-th** success in independent repetitions of a Bernoulli trial with probability p of success at every trial. If X denotes this r.v., can you list the values taken by X ? X takes values r, r + 1, . . . We wish to obtain P[X = j] for $j \geq r$.

The event [X = j] occurs iff there are (r – 1) successes in the first (j – 1) trials and the j-th trial results in a success. In view of independence of the successive trials,

P[X = j] = P[There are (r – 1) successes in the first (j – 1) trials and the j-th trial results in a success]

= P[There are (r – 1) successes in the first (j – 1) trials] × P[The j-th trial results in a success].

Now, recall the argument which we used to find the probabilities related to the binomial distribution (Sec. 8.3). By a similar argument we get

P[There are (r – 1) successes in the first j – 1 trials]

$$= \binom{j-1}{r-1} p^{r-1} (1-p)^{(j-1)-(r-1)}$$

$$= \binom{j-1}{r-1} p^{r-1} (1-p)^{j-r}, \; j \geq r.$$

Moreover,

P[jth trial results in a success] = p.

Hence,

$$P[X = j] = \binom{j-1}{r-1} p^{r-1} (1-p)^{j-r} p$$

$$= \binom{j-1}{r-1} p^{r} (1-p)^{j-r}, \; j = r, r+1, \ldots$$

This leads us to the following definition.

Definition 2 : A r.v. X has **negative binomial distribution** with parameters (r, p), r a positive integer and $0 < p < 1$, if the p.m.f. of X is given by

$$f(j; r, p) = P[X = j] = \binom{j-1}{r-1} p^{r} (1-p)^{j-r}, \; j = r, r+1, \qquad \ldots (4)$$

Now let us verify that

$$\sum_{j=r}^{\infty} f(j; r, p) = \sum_{j=r}^{\infty} \binom{j-1}{r-1} p^{r} (1-p)^{j-r} = 1$$

for all positive integral r and $0 < p < 1$. We do this in Theorem 3. But before that we need some preparation.

Recall that the symbol $\binom{n}{j}$ stands for the number of ways of choosing j objects out of n distinct objects. Here n is a positive integer and j is a non-negative integer. We want to extend the definition of $\binom{n}{j}$ when n is replaced by any real number α, say, $-\infty < \alpha < \infty$. You know that

$$\binom{n}{j} = \frac{\{n(n-1)\ldots(n-j+1)\}}{j!} \qquad \ldots (5)$$

You will agree that the right side of (5) makes sense even if n is not a positive integer. We therefore define

$$\binom{\alpha}{j} = \frac{\alpha(\alpha-1)\ldots(\alpha-j+1)}{j!} \qquad \ldots(6)$$

$-\infty < \alpha < \infty$, j being a non-negative integer.

The advantage of this extension is that we can write down the expansion,

$$(1+t)^{\alpha} = 1 + \binom{\alpha}{1}t + \binom{\alpha}{2}t^2 + \ldots, \qquad \ldots(7)$$

which is valid for all real α and $-1 < t < 1$. Formula (7) is known as **Newton's binomial formula.**

If α is a positive integer n, the right side of (7) consists of (n + 1) terms, since $\binom{n}{j}$ is zero for $j > n$. In fact, in this case, (7) is the usual binomial expansion of $(1+t)^n$ and is valid for all real t.

If α is not a positive integer, the right side of (7) is an infinite series which is convergent only for $-1 < t < 1$.

Now we first note that

$$\binom{j-1}{r-1} = \binom{j-1}{j-1-(r-1)} = \binom{j-1}{j-r}, \; j = r, r+1, \ldots$$

In this relation put $k = j - r$, so that $k = 0, 1, 2 \ldots$ and we have

$$\binom{j-1}{r-1} = \binom{r+k-1}{k}$$

$$= \frac{(r+k-1)(r+k-2)\ldots(r+k-1-k+1)}{k!}$$

$$= \frac{r(r+1)\ldots(r+k-1)}{k!}$$

$$= (-1)^k \frac{(-r)(-r-1)\ldots(-r-k+1)}{k!},$$

writing the terms in the numerator in the reverse order.

$$= (-1)^k \binom{-r}{k} \qquad \ldots(8)$$

We shall use this result in the proof of the following theorem.

Theorem 3 : The sum of the negative binomial probabilities f(j; r, p) is one, i.e.

$$\sum_{j=r}^{\infty} \binom{j-1}{r-1} p^r (1-p)^{j-r} = 1.$$

Proof : Write $q = 1 - p$ and $j - r = k$. Then using (8) we get

$$\sum_{j=r}^{\infty} \binom{j-1}{r-1} p^r (1-p)^{j-r} = p^r \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} q^k$$

$$= p^r \sum_{k=0}^{\infty} \binom{-r}{k} (-q)^k$$

$$= p^r (1-q)^{-r}, \qquad \text{using (7)}$$

$$= 1, \text{ since } 1 - q = p.$$

The above discussion also brings out the fact that the negative binomial probabilities f(j; r, p), $j \geq$ are terms of the binomial expansion of $p^r (1-q)^{-r}$, which has a negative exponent, (–r). It is for this reason that the probability distribution specified by (4) is called the negative binomial distribution. It is also known as the **Pascal distribution.**

Blaise Pascal (1623–1662)

In the following examples you will see some situations where the r.v. has negative binomial distribution.

Example 3: A proof-reader catches a misprint with probability 0.60. Let us find the probability that a total of ten misprints have occurred before our proof-reader catches his third misprint.

If our proof-reader catches a misprint, we'll term it a success! Here we want to find the probability that the third success occurs at the tenth trial, when p = 0.60. Hence with r = 3, and j = 10, the required probability is

$$f(10; 3, 0.60) = \binom{9}{2}(0.60)^3 (0.40)^7 = 0.013.$$

Example 4: The probabilities of having a male or a female child are both 0.50. Can you find the probability that a family's fourth child is their second daughter?

Let us term the birth of daughter a success.

We have p = 1/2, and we need

$$f(4; 2, 0.5) = \binom{3}{1}(0.5)^2 (0.5)^2$$

$$= \frac{3}{16}.$$

In the following discussion we evaluate the mean and variance of the negative binomial distribution with parameters (r, p).

Notice that the number X of trials required for the rth success is the sum of r r.vs., $Y_1, Y_2, \ldots, Y_r$, where Y_1 is the number of trials required for the first success, Y_2 is the number of trials required, after the first success, to obtain the second success, and so on. In general, Y_j is the number of trials between the (j – 1)th and j-th success. Do you agree that $Y_1, Y_2, \ldots, Y_r$ are independent r.vs. and that each has the geometric distribution with the same parameter p?

It follows from Theorem 1, that

$$E(Y_j) = \frac{1}{p}, \quad \text{Var}(Y_j) = \frac{1-p}{p^2}.$$

Hence,

$$E(X) = E(Y_1 + Y_2 + \ldots + Y_r) = \frac{r}{p}, \text{ and} \qquad \ldots (10)$$

$$\text{Var}(X) = \text{Var}(Y_1 + \ldots + Y_r)$$

$$= \sum_{j=1}^{r} \text{Var}(Y_j) = \frac{r(1-p)}{p^2}. \qquad \ldots (11)$$

Caution : The above discussion only indicates a method of derivation of E(X) and Var(X), and is not a formal proof of (10) and (11)

We are sure you will be able to solve the following exercises on the basis of our discussion in this section.

E3) Find the probability that a person tossing an unbiased coin gets fourth head on seventh toss.

E4) Find the probability that a person rolling an unbiased die, gets his third six on the eighth roll.

E5) A scientist innoculates several mice, one at a time, with a virus which produces a disease in them. If each mouse has probability 1/4 of developing the disease, find the expected number of mice required for an experiment in which the scientist stops after obtaining the second mouse with the disease.

E6) Compute the moment generating function of the negative binomial distribution.

E7) Let X and Y be two independent r.vs. with negative binomial distributions and parameters (r, p) and (s, p), respectively. Find the m.g.f. of X + Y.

So far, we have seen that the geometric distribution can be applied to situations where we are interested in the number of trials needed for the **first success.** On the other hand, the negative binomial distribution applies to situations in which our interest lies in the number of trials required for r **successes**, where r is a positive integer. So what happens if we take r = 1 in the negative binomial distribution? We get the geometric distribution, of course.

In the next section we take up one last discrete probability distribution—the Poisson distribution.

9.4 THE POISSON DISTRIBUTION

We describe below three real-life situations from three different areas. The first case is from meteorology in which we are concerned with the frequency with which rain storms occur. The second case is related to frequency of wrong telephone connections and the third is related to bacterial counts in different areas of dish called the Petri plate which biologists use. We shall then describe their common features. These can be used to develop a probability distribution, called the Poisson distribution, in honour of the French mathematician Simeon D. Poisson (1781-1840) who studied it for the first time.

Simeon D. Poisson (1781–1840)

Case 1 : The table below is based on the records of 10 rainfall stations over a period of 33 years. Thus we have records for $10 \times 33 = 330$ station-years. This table gives the number of rainstorms, i.e. the number of 10 minute periods with more than 1 cm. of rain.

Table 1 : Rainstorms

x	0	1	2	3	4	5
Frequency	102	114	74	28	10	2

Source : E.L. Grant (1964), Statistical Quality Control.

Here x is the number of rainstorms in a station-year and the corresponding frequency is the number of station-years with x rainstorms.

Case 2 : Table 2 shows the frequency distribution of telephone connections to a wrong number. A total of 267 telephones were observed.

Table 2: Connections to wrong numbers

x	0-2	3	4	5	6	7	8	9	10
Frequency	1	5	11	14	22	43	31	40	35

x	11	12	13	14	15	More than 16
Frequency	20	18	12	7	6	2

Source: W.Feller (1972), An Introduction to Probability Theory and its Applications, Vol. I.

Here x is the number of wrong telephone connections and the frequency gives the number of telephones with x wrong connections.

The concept of a station-year is similar to that of man-hour. If three men work for 8 hours each, we say that they have worked for $3 \times 8 = 24$ man-hours.

Case 3 : Bacterial colonies develop over the surface of a Petri plate. The plate is divided into a large number of small squares of equal area and observed under a microscope. The bacterial colonies are visible as dark spots. The following table gives the observed number (frequency) of squares with exactly x dark spots.

Table 3 : Bacterial Counts

x	0	1	2	3	4	5	6 or more
Frequency	5	19	26	26	21	13	8

Source : W. Feller (1972), An Introduction to Probability Theory and its Applications, Vol. I.

On the face of it, there is very little similarity between these three cases. However, notice that in each case we have counted the number of times an event has occurred. The event concerned has many opportunities or trials when it could have occurred but it had a very small probability of occurrence at given trial. Thus,

- there are many 10 minute periods in a year, buit it is very unlikely that any specific 10-minute interval would have a rainstorm.
- there are many occasions when any one of the 267 telephones would be used but the chance of a wrong connection can be expected to be small.
- a Petri plate has a large number of small squares and it would be rare to find a bacterial colony in a specified square.

In other words, we can think of a large number, n, of independent Bernoulli trials with a small probability p of 'success' at each trial. Although n is large and p is small, we can expect the mean number np of successes to be a finite number. Thus, we are interested in the probability distribution of the number of successes in a large number n of independent Bernoulli trials, each with the same small chance p of success such that np remains finite. We know that the number of successes in n such independent trials follows a binomial distribution. So, the probability b(r; n, p) of r successes in n independent Bernoulli trials with constant probability p of success is

$$b(r; n, p) = \binom{n}{r} p^r (1-p)^{n-r}, \; r = 0, 1, \ldots, n.$$

But we want to find what happens when n is large and p is small. That is, we want to find the limit of b(r; n, p) as $n \to \infty$ and $p \to 0$, such that np equals m, say, where m is a positive number.

We can do this as follows :

We have p = m/n, and

$$\begin{aligned} b(r; n, p) &= \binom{n}{r} p^r (1-p)^{n-r} \\ &= \frac{n(n-1)\ldots(n-r+1)}{r!} (m/n)^r (1-m/n)^{n-r} \\ &= \frac{1\,(1-1/n)\,(1-2/n)\cdot(1-(r-1)/n)}{r!} m^r (1-m/n)^{n-r} \end{aligned}$$

The factor $1 \cdot (1-1/n) \ldots [1-(r-1)/n]$ converges to 1 as $n \to \infty$. Moreover, the term

$$(1-m/n)^{n-r} = \frac{(1-m/n)^n}{(1-m/n)^r} \to \frac{e^{-m}}{1} = e^{-m}$$

as $n \to \infty$, r being kept fixed. The conclusion is that

$$b(r; n, p) \to \frac{e^{-m} m^r}{r!} = p(r, m), \text{ say} \qquad \ldots (12)$$

as $n \to \infty$, $p \to 0$, such that np = m.

There are two ways of looking at (12).

- One is to treat p(r, m) as an approximation to b(r; n, p). In fact, we call p(r, m), the **Poisson approximation to b (r; n, p).**
- Another way is to regard

$$p(r, m) = \frac{e^{-m} m^r}{r!}, \; r = 0, 1, 2, \ldots \qquad \ldots (13)$$

as the p.m.f. of a r.v. It is easy to verify that p(r, m) has all the qualifications to be a p.m.f., since

$$p(r, m) > 0 \text{ for all } r = 0, 1, 2, \ldots$$

$$\text{and } \sum_{r=0}^{\infty} p(r, m) = e^{-m} \sum_{r=0}^{\infty} m^r/r! = e^{-m} e^m = 1.$$

In this case we give the following definition.

Definition 3 : A r.v. X is said to have **Poisson distribution** with parameter m > 0, if its p.m.f. is

$$p(r, m) = P[X = r] = \frac{e^{-m} m^r}{r!}, \; r = 0, 1, 2, \ldots \qquad \ldots (14)$$

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = \lim_{t\to\infty}\left(1+\frac{1}{t}\right)^{-mt}$$

$$= \left[\lim_{t\to\infty}\left(1+\frac{1}{t}\right)^t\right]^{-m}$$

$$= e^{-m}$$

since $\lim\left(1-\frac{1}{t}\right)^t = e$

(Recall Unit 5, MTE-01).

Now let us compare the probabilities obtained by applying the binomial and Poisson distributions with the help of an example.

Example 5 : There are few printing mistakes in the material printed at a good press. In fact, the probability of a printing mistake is 0.01. Let us find the probability that in a text with 500 words, there are no mistakes.

Assuming that the conditions for binomial distribution hold, the required probability is

$$b(0; 500, 0.01) = (0.99)^{500} \simeq 0.0066.$$

Suppose we use the Poisson approximation for b(0; 500, 0, 0.1). Since n = 500, p = 0.01, we may take m = np = 5. Hence,

$$p(0, 5) = e^{-5} = 0.0067.$$

Notice that the difference is only in the fourth place of decimal.

The following table gives the values of b(r, 500, 0.01) for r = 0, 1, 2, 3, 4 and those of the corresponding Poisson approximations p(r, 5), for the same values of r.

Table 4 : Probability of r printing mistakes

r	0	1	2	3	4
Binomial distribution f(r; 500, 0.01)	0.0066	0.0335	0.0840	0.1408	0.1768
Poisson approximation p(r, 5)	0.0067	0.0335	0.0838	0.1396	0.1745

You would notice that the Poisson approximation is quite satisfactory. In fact, it would improve with larger values of n and smaller values of p. Generally speaking, the Poisson approximation to the binomial probabilities is satisfactory if $n \geq 20$ and $p \leq 0.05$.

In calculating the above probabilities we have used the recurrence relation for binomial probabilities. We have also used the following recurrence relation for the Poisson probabilities.

We have

$$p(r+1, m) = \frac{e^{-m} m^{r+1}}{(r+1)!}$$

$$= \frac{\{e^{-m} m^r / r!\}}{(r+1)}$$

$$= \frac{mp(r, m)}{r+1},$$

which is valid for all r = 0, 1, 2, . . .

Let us now obtain the mean and variance of the Poisson distribution.

Theorem 4 : If the r.v. X has Poisson distribution with parameter m, then

$$E(X) = m, \quad Var(X) = m.$$

Proof : We have by definition

$$E(X) = \sum_{r=0}^{\infty} r\, P[X = r]$$

$$= \sum_{r=0}^{\infty} r\, e^{-m} \frac{m^r}{r!}$$

$$= m\, e^{-m} \sum_{r=1}^{\infty} \frac{m^{r-1}}{(r-1)!}$$

$$= m\, e^{-m} \sum_{t=0}^{\infty} \frac{m^t}{t!}, \text{ where } t = r - 1.$$

Since $\sum_{r=0}^{\infty} \frac{m^t}{t!} = e^m$, it follows that

$$E(X) = me^{-m} e^m = m.$$

The first step in the calculation of the variance is to compute

$$E\{X(X-1)\} = \sum_{r=0}^{\infty} r(r-1)\, P[X=r]$$

$$= \sum_{r=2}^{\infty} r(r-1)\, P[X=r]$$

$$= \sum_{r=2}^{\infty} r(r-1)\, e^{-m} \frac{m^r}{r!}$$

$$= m^2 e^{-m} \sum_{r=2}^{\infty} \frac{m^{r-2}}{(r-2)!}$$

$$= m^2 e^{-m} \sum_{u=0}^{\infty} \frac{m^u}{u!} \quad , \text{where } u = r-2$$

$$= m^2 e^{-m} e^m = m^2$$

Now recall that

$$\text{Var}(X) = E\{X(X-1)\} + E(X) - \{E(X)\}^2$$

$$= m^2 + m - m^2 = m.$$

Thus, the results of the theorem are established.

So, the mean and variance of the Poisson distribution are always equal.

The next theorem gives us the m.g.f.

Theorem 5 : The moment generating function of the Poisson distribution is

$$M_x(t) = \exp\{m(e^t - 1)\}.$$

valid for all real t.

Proof : We have

$$M_x(t) = E[e^{tX}]$$

$$= \sum_{r=0}^{\infty} e^{tr} e^{-m} \frac{m^r}{r!}$$

$$= e^{-m} \sum_{r=0}^{\infty} \frac{(me^t)^r}{r!}$$

$$= \exp\{-m + me^t\}$$

$$= \exp\{m(e^t - 1)\},$$

as required.

We can use this theorem to prove the additive property of variables with Poisson distribution.

Corollary : If X_1 and X_2 are independent Poisson r.vs with parameters m_1 and m_2, respectively, then $X_1 + X_2$ has Poisson distribution with parameter $m_1 + m_2$.

Proof : We first determine the probability,

$P[X_1 + X_2 = k]$. The event $X_1 + X_2 = k$ is the union of the mutually exclusive events, $X_1 = 0, X_2 = k, X_1 = 1, X_2 = k - 1, \ldots ; X_1 = k, X_2 = 0$.

Therefore,

$$P[X_1 + X_2 = k] = \sum_{j=0}^{k} P[X_1 = j,\ X_2 = k - j]$$

$$= \sum_{j=0}^{k} P[X_1 = j]\ P[X_2 = k - j]$$

$$= \sum_{j=0}^{k} \frac{e^{-m_1} m_1^j}{j!} \frac{e^{-m_2} m_2^{k-j}}{(k-j)!}$$

$$= \frac{e^{-(m_1 + m_2)}}{k!} \sum_{j=0}^{k} \binom{k}{j} m_1^j\ m_2^{k-j}$$

$$= \frac{e^{-(m_1 + m_2)}}{k!} (m_1 + m_2)^k.$$

This shows that the p.m.f. of $X_1 + X_2$ is that of a Poisson r.v. with parameter $(m_1 + m_2)$ and hence, $X_1 + X_2$ has a Poisson distribution with parameter $(m_1 + m_2)$.

We can easily extend this result to more than two variables.

Corollary : If $X_1, X_2, \ldots, X_n$ are independent Poisson variates with parameters $m_1, m_2, \ldots, m_n$, respectively, then the r.v. $X_1 + X_2 + \ldots + X_n$ has Poisson distribution with parameter $m_1 + m_2 + \ldots + m_n$.

We have seen that Poisson distribution gives a very good approximation of binomial distribution. Poisson distribution can also arise in situations which have no direct connection with the binomial distribution. But we shall not discuss such situations here.

See if you can solve these exercises now.

E8) Records show that the probability that a train has an accident between two specific stations is 0.0004. Use the Poisson approximation to the binomial probabilities to obtain the probability that in its 700 trips during the year, the train would have at most one accident.

E9) It is known that the number of imperfections per metre of a certain variety of cloth is a Poisson r.v. with m = 0.12. Find the probability that ten metres of this cloth will have

a) four imperfections

b) at most three imperfections.

Hint : Use the second corollary to Theorem 5, assuming that imperfections over non-overlapping portions of the cloth are independent.

E10) a) Compute the mean, m, for the frequency distribution given in Table 1.

b) Use this value of m to calculate the Poisson probabilities of 0, 1, . . . , 6 rainstorms.

c) The product $N \times p(r, m)$, where N = 330 is the total number of observations, gives the expected frequency based on the assumption that the number of rainstorms occur according to Poisson distribution, Fill in the blanks in the following table :

Table 5 : Number of Rainstorms

x	0	1	2	3	4	5	6
Observed frequency	102	114	74	28	10	2	0
Expected frequency							

E11) Fill in the blanks in the following tables on the assumptions tahat the variables have Poisson distribution.

a)

Table 6 : Number of wrong connections

x	Observed frequency	Expected frequency
0-2	1	
3	5	
4	11	
5	14	
6	22	
7	43	
8	31	
9	40	
10	35	
11	20	
12	18	
13	12	
14	7	
15	6	
≥ 16	2	
Total	267	

b)

Table 7 : Counts of bacteria

x	0	1	2	3	4	5	> 6
Observed frequency	5	19	26	26	21	13	8
Expected frequency							

When you have done E10 and E11, you will find that the agreement between observed and expected frequencies is quite good in all the three cases. You will study the methods of comparing the observed and expected frequencies in more detail in Block 4 under the topic "chi-square tests of goodness of fit".

Now let us summarise what we have done in this unit.

9.5 SUMMARY

In this unit we have covered the following main points :

1) The geometric distribution and the negative binomial distribution are two examples of waiting time distributions. They are the distributions of the number of trials required for the first and the r-th success in independent repetitions of Bernoulli trials. Thus, the geometric distribution is a particular case (when $r = 1$) of the negative binomial distribution. Moreover, the negative binomial distribution can be regarded as the distribution of the sum of r independent and identically distributed geometric r.vs. with parameter p.

2) The Poisson distribution is in a different class. It can be regarded as the limiting form of a binomial distribution obtained by allowing $n \to \infty$ and $p \to 0$ such that np is finite. This approach enables us to compute approximately the binomial probabilities.

3) The negative binomial and the Poisson distribution also possess the so-called reproductive property :

   If X_1 and X_2 are independent r.vs having negative binomial distributions (Poisson distributions) with parameters (r_1, p) and (r_2, p) (m_1 and m_2), respectively,

then $X_1 + X_2$ also has the negative binomial (Poisson) distribution with parameters $(r_1 + r_2, p)$ $(m_1 + m_2)$.

The standard distributions which we described in Units 8 and 9 are not the only discrete distributions. There are many others with interesting properties which we hope you would feel inclined to study in the future.

9.6 SOLUTIONS AND ANSWERS

E1) We need to obtain $P[X \geq 6]$ when X has the geometric distribution with $p = 1/6$. The required probability is

$$P[X \geq 6] = \sum_{r=6}^{\infty} P[X = r]$$

$$= \sum_{r=6}^{\infty} (1/6)(5/6)^{r-1}$$

$$= (5/6)^5 \simeq 0.402.$$

E2) The required probability is

$$1/13 + (1/13)(12/13) + (1/13)(12/13)^2 + (1/13)(12/13)^3 \simeq 0.269.$$

E3) We need

$$f(7; 4, 1/2) = 0.1556.$$

E4) The required probability is

$$f(8; 3, 1/6) = 0.039.$$

E5) We need E(X) when X has negative binomial distribution with $r = 2$, $p = 1/4$. The answer is $E(X) = 8$.

E6) If X has the negative binomial distribution with parameters r and p, its m.g.f. is

$$M_x(t) = \sum_{j=r}^{\infty} e^{jt} \binom{j-1}{r-1} p^r q^{j-r}$$

$$= p^r e^{rt} \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} (qe^t)^k$$

$$= \frac{p^r e^{rt}}{\{1 - qe^t\}^r},$$

Where $q = 1 - p$ and $t < \ln\{(1-p)^{-1}\}$.

E7) The m.g.f. of X + Y is

$$p^{r+s} \exp\{(r+s)t\} / (1 - qe^t)^{r+s},$$

provided $t < \ln\{(1-p)^{-1}\}$.

E8) We have $m = 0.0004 \times 700 = 0.28$ and we need

$$p(0, 0.28) + p(1, 0.28) \simeq 0.967.$$

E9) In view of the corollary, the number X of imperfections in ten metres of the cloth has Poisson distribution with

$$m = 10 \times 0.12 = 1.2.$$

a) $p(4, 0.12) \simeq 0.026$

b) $p(0, 1.2) + p(1, 1.2) + p(2, 1.2) + p(3, 1.2) \simeq 0.966.$

E10) a) $m = 1.2$

b)

Number rainstorms	0	1	2	3	4	5	6
Probability	0.3	0.36	0.21	0.088	0.027	0.006	0.003

c)

Number of Rainstorms

x	0	1	2	3	4	5	6
Observed frequency	102	114	74	28	10	2	0
Expected frequency	99	119	71	29	9	2	1

E11) a)

Number of wrong connections

x	Observed frequency	Expected frequency
0-2	1	2.05
3	5	4.76
4	11	10.39
5	14	18.16
6	22	26.45
7	43	33.03
8	31	36.01
9	40	35.04
10	35	30.63
11	20	24.34
12	18	17.72
13	12	11.92
14	7	7.44
15	6	4.33
≥ 16	2	4.65
Total	267	267.00

b)

Counts of bacteria

x	0	1	2	3	4	5	> 6
Observed frequency	5	19	26	26	21	13	8
Expected frequency	6.1	180	26.7	26.4	19.6	11.7	9.5

NOTES

NOTES

UTTAR PRADESH
RAJARSHI TANDON OPEN UNIVERSITY

UGMM - 11
Probability and Statistics

Block

3

DISTRIBUTION THEORY

Course Design Committee

Prof. S.K. Mitra (*Chairman*)
Indian Statistical Institute
New Delhi

Prof. A.M. Goon
Presidency College
Calcutta

Prof. J. Medhi
Guwahati

Prof. B.L.S. Prakasa Rao
Indian Statistical Institute
New Delhi

Prof. Aloke Dey
Indian Statistical Institute
New Delhi

Prof. K. Balasubramanian
Indian Statistical Institute
New Delhi

Prof. D.D. Joshi
Ex-Pro-Vice-Chancellor
IGNOU

Dr. V. Madan
School of Sciences
IGNOU

Dr. Poornima Mital
School of Sciences
IGNOU

Dr. Manik Patwardhan
School of Sciences
IGNOU

Dr. Sujatha Varma
School of Sciences
IGNOU

Block Preparation Team

Prof. S.K. Mitra (*Editor*)
ISI, New Delhi

Prof. Aloke Dey (*Co-editor*)
ISI, New Delhi

Prof. B.L.S. Prakasa Rao
Indian Statistical Institute
New Delhi

Dr. Manik Patwardhan
School of Sciences
IGNOU

Dr. Sujatha Varma
School of Sciences
IGNOU

Course Coordinator : Prof. R.K. Bose

Production

Mr. Balakrishna Selvaraj
Registrar (PPD)
IGNOU

January, 1994

ISBN - 81-7263-532-x

Reprinted by : Nitin Printers, 1 Old Katra, Manmohan Park, Allahabad.

BLOCK 3 DISTRIBUTION THEORY

In Block 2, we have discussed how to build probability models and the corresponding probability distributions on discrete sample spaces. In this block, we extend the concept of probability and probability distributions to sample spaces with the number of elementary outcomes possibly uncountable.

In Unit 10, we define the notion of a random variable and its distribution function and study their properties. Some important examples of univariate distributions, which are commonly used, are discussed in Unit 11. The notion of joint distributions of a random variable or bivariate distribution is introduced in Unit 12. Related notions of independence, covariance and correlation are also studied in this unit. Material discussed in Units 10 to 12 will help you in understanding the basic concepts involved in studying the relation between different characteristics.

Unit 13 deals with distributions of a function of a random variable. Some important distributions which are distributions of functions of standard normal random variables are discussed in this unit. Results obtained in this unit will help you in the study of statistical methods for data analysis in Block 4. Finally, we discuss the law of large numbers and the central limit theorem in Unit 14.

In the next block we shall study statistical methods for analysing a data. But we shall often use what is covered in this block. So, before going to the next block, please ensure that you have achieved the objectives of the units in this block.

Notations and Symbols

F	:	Distribution function
f	:	Density function
Φ	:	Standard normal distribution
φ	:	Standard normal density
$F_{X,Y}$	:	Distribution function of the bivariate random vector (X, Y)
$f_{X,Y}$	:	Joint probability density function
F_X	:	Marginal distribution of X
f_X	:	Marginal density of X
$f_{X\|Y}$	:	The conditional density function of X given Y = y
$f_{Y\|X}$	:	The conditional density function of Y given X = x
$F_{X\|Y}$	:	The conditional distribution of X given Y = y
$F_{Y\|X}$	:	The conditional distribution of Y given X = x
$N(\mu\ \sigma^2)$:		The normal distribution with mean μ and variance σ^2.

Also see the lists in Blocks 1 and 2.

Acknowledgement

To Prof. R.K. Bose, Dr. Parvin Sinclair, for their useful comments on the manuscript.

UNIT 10 UNIVARIATE DISTRIBUTIONS

Structure

10.1 INTRODUCTION

In this unit, we first introduce the concept of a distribution function of a random variable. Random variables taking values in either a finite set or countably infinite set have been studied in Unit 6. Our main emphasis here is on random variables taking values in a set which is possibly uncountable. Most often, we consider random variables where values fall in an interval, finite or infinite, on the real line. A special class of distributions, namely, absolutely continuous distribution play a major role in practical problems. Throughout this unit, this class of distributions is the base of our study. We will discuss the notions of a function, expectation and the variance of a random variable in Secs. 10.3–10.5. The concept of moment of a random variable and a method of obtaining moments using the moment generating function are given in Sec. 10.6 and Sec. 10.7. Different approaches useful in finding the probability distribution function of functions of a given random variable are discussed in Sec. 10.8.

The facts covered in this unit will be used constantly in the rest of the course. Therefore we suggest that you do all the exercises in the unit as you come to them. We will use some facts from Blocks 1, 2, 3 and 4 of MTE–01 and Block 1 of MTE–07. So keep them handy while studying, so that you can refer to them easily. Further, please do not go to the next unit till you are sure that you have achieved the following objectives.

Objectives

After reading this unit you should be able to :

- define the distribution function for a random variable and a density function for an absolutely continuous distribution, and establish their interrelations ;
- check whether a given function is a distribution function ;
- check whether a given function is a density function ;
- compute the distribution function and the density function when it exists ;
- compute the moments and moment generating function of a random variable when they exist ;
- derive the distribution function of a function of a random variable.

10.2 DISTRIBUTION FUNCTIONS

In Block 2, we have discussed the concept of probability on discrete sample spaces at length. If you remember, we had started our discussion with the definitions of random experiments and their sample spaces. We had then remarked that sample spaces can be classified as discrete and continuous. Since the treatment for these two categories is slightly different, we had then focussed our attention only on the discrete case. Now we take up the case of general spaces and define probabilities. Where do we begin ? As before corresponding to a random experiment, we have a sample space. We call each of its elements (sample points) an **outcome** or an **elementary event.** But what about an event ?

In a discrete sample space S, we say that **any** subset of S is an event. In other words, the collection of all subsets of S is precisely the collection of all events. Now, in a general sample space Ω, it is not always possible to consider all subsets of Ω as events. There are some difficulties in doing this which we shall not explain to you as the technicalities are beyond the level of this course. So we are forced to take a smaller collection of subsets of the sample space as the collection of all events. But, at the same time, we would like this collection of events to have certain "reasonable" properties. For example, we would like

i) Ω to be an event.

ii) If A is an event, then A^c should also be an event.

iii) If $A_1, A_2, \ldots$ are events, then $\bigcup_{i=1}^{\infty} A_i$ should also be an event.

We are sure you will have no problem in agreeing to properties (i) and (ii) above. What about the third one ? If there are only a finite number of events in Ω, then even this property seems reasonable. (In fact, you have already come across it in Block 2.) At this stage, we can only say that the third condition is an important axiom, which is crucial to the development of the probability concept. An important point to note here is that in (iii) above, we are taking only countably infinite unions and not uncountably infinite unions.

To take into account the properties (i), (ii) and (iii) above, we define a collection $\mathcal{F}$ of subsets of the sample space Ω, which has the following properties :

i) $\Omega \in \mathcal{F}$

ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

iii) If each of $A_1, A_2, \ldots$ belong to $\mathcal{F}$, then

$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

σ is the greek letter 'sigma'.

Remark 1 : Note that property (iii) guarantees only that the union of a **countable** number of sets belongs to $\mathcal{F}$. It does not say anything about the union of an uncountable number of sets. The above collection $\mathcal{F}$ is called a σ – **field of events** of Ω.

We say that A is an **event** in the sample space Ω if $A \in \mathcal{F}$.

Now that we have defined events, let us talk about their probabilities. In the discrete case we had associated probabilities to each outcome and then added these up to calculate probabilities of events. But it is not always possible to do this in general (not necessarily discrete) sample spaces. We can give you a glimpse of the kind of difficulties that we may encounter.

Let us consider the random experiment of choosing a number x, at random, from the interval [0, 1]. This means that the probability assigned to each value in [0, 1] should be the same. But the total probability assigned to [0, 1] is one. This leads us to assign a probability zero to each individual value in [0, 1]. If the aggregate of the

probability of each individual value x, $x \in [0, 1]$ is taken to be the probability of [0, 1], (if such an aggregation in the sense of summation of individual values is possible) then we encounter a problem in that although $P\{[0, 1]\} = 1$, the individual terms in the aggregate are all zero. In fact the same difficulty would be faced with a discrete sample space having countably infinite sample points, for example when one desires to draw an integer at random with equal probability from the set of all positive integers. This raises the question : What do we mean by an aggregate of an uncountable number of values ? Is such an aggregation at all possible ?

We take care of this in the following definition.

Definition 1 : Let P be a real-valued function defined on $\mathcal{F}$, the collection of events on a sample space Ω. Suppose P has the following properties :

i) $P(\Omega)=1$

ii) $0 \leq P(E) \leq 1 \; \forall \; E \in \mathcal{F}$

iii) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ if $A_i \in \mathcal{F}$ for $i \geq 1$ and $A_i \cap A_j = \phi$ for $i \neq j$.

Then, P is called a **probability function.**

The problem that we encountered when we were taking the aggregate of individual probabilities to obtain the probability of the union is bypassed by the above definition because we have conveniently disregarded uncountable unions in (iii) of the above definition. As such, with this definition of probability, it becomes meaningless to talk about an aggregation of probabilities over an uncountable set.

From Definition 1 we have $P(\Omega) = 1$. Can you deduce the value of $P(\phi)$ from this ? Do you agree that $P(\phi) = 0$? Suppose $P(\phi) \neq 0$, then it equals some positive number, say r, in]0, 1].

Now $\Omega = \Omega \cup \phi$ and $\Omega \cap \phi = \phi$.

$\therefore$ By (iii) in Definition 1,

$P(\Omega) = P(\Omega)+P(\phi) = 1+r > 1$.

This is a contradiction. Hence $P(\phi) = 0$.

In Unit 5 we had listed some examples of sample spaces which are not discrete. You might have noticed that in each of these examples, the outcome is expressed in numerical terms. In most other practical situations as well, we can assign a real number to each outcome in the continuous sample space. This observation allows us to consider only those sample spaces which are subsets of **R**, the set of real numbers.

Now, let us consider a continuous sample space S. We saw that we can associate a real number to each outcome of S. Does this correspondence have any significance ? Before answering this question let us go back for a moment to the discrete case. Recall from Unit 7 (Block 2) that, if the sample space is discrete then we can associate a number to each outcome, and this association defines a real-valued function on the discrete sample space. This function is called a discrete random variable.

If X denotes a random variable taking values $x_1, x_2, \ldots$, then the probability mass function is defined by $f(x_j) = P[X = x_j]$.

We have also seen in Unit 7 that the importance of a random variable lies in the fact that using that we can define another function called probability mass function : Thus the probability mass function gives the probability of occurrence of the elements in the range of X which in turn can be used to compute the probability of occurrence of any event defined by the observed values of X.

Now can we define a continuous random variable in the same way as in the discrete case ? From the preceding discussion we know that the definition of a random variable should conform with the definition of probability mass function. But in the continuous case, there are some constraint induced on what kind of subsets of Ω can

be assigned a probability. This imposes certain conditions on the definition of random variables. More clearly, suppose X is a random variable (r.v.) and we want to evaluate the probability that the random variable X takes values in a set $A \subseteq \mathbf{R}$, i.e. $P[X \in A]$. Then we are actually concerned with the set $B = \{\omega: X(\omega) \in A\} \subseteq \Omega$, and we want to evaluate the probability of this subset of Now we know that we can obtain the probability if B only if B belongs to the speci class $\mathcal{F}$ of sets we defined earlier. So naturally, we need to modify the definition of a random variable. We, thus have the following definition of a r.v.

Definition 2 : Let ε be an experiment and Ω, the sample space associated with it. L $\mathcal{F}$ be the collection of events in Ω. A real-valued function X, defined on Ω is called a **random variable** (r.v.), if

$$[X \leq x] = \{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F} \forall\ x \in \mathbf{R}.$$

If we study one such real-valued function defined on Ω, we have a univariate problem under study. If we simultaneously study two such real-valued function on Ω, we have a bivariate problem and so on. Bivariate distributions will be studied in Unit 12.

Next we shall define another function related to random variables which can be use to evaluate probabilities of events.

Definition 3 : The distribution function F for a random variable X is a function defined on the real-line by

$$F(x) = P[X \leq x]$$

where $-\infty < x < \infty$.

The definition makes sense because if X is a random variable, then $[X \leq x]$ is an event in Ω. Therefore $P[X \leq x]$ is well-defined. This function is sometimes called t **cumulative univariate distribution function.** You know why this is called univariate, isn't it ? This is because the corresponding random variable is one variable. Our discussion from now on deals with random variables and their distributions. So you won't have to worry about the nature of the σ–field $\mathcal{F}$.

Let us now try to understand the distribution function by looking at some of its properties.

Properties of a distribution function F(x)

a) $0 \leq F(x) \leq 1$ for all $x \in \mathbf{R}$.

This property is a consequence of the property (ii) of the probability function since every event $[X \leq x]$ should have a number between 0 and 1 as its probability.

b) **F(x) is a non-decreasing function of** x ; that is, if $x \leq y$, then $F(x) \leq F(y)$.

To obtain this property, we write

$$[X \leq y] = [X \leq x] \cup [x < X \leq y].$$

Since the events $[X \leq x]$ and $[x < X \leq y]$ are disjoint, by the property (iii) of the probability function P, we have

$$P[X \leq y] = P[X \leq x] + P[x < X \leq y].$$

But by the property (ii) of the probability function, the last term, namely, $P[x < X \leq y] \geq 0$. Hence we get

$$P[X \leq y] \geq P[X \leq x].$$

That is,

$$F(y) \geq F(x).$$

Did you notice that the above argument also proves that

$P[x < X \leq y] = F(y) - F(x)$ if $x < y$?

Next we shall state two more properties. We have omitted the verifications of these properties as they are too technical.

c) $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.

We also write Property (c) in the form $F(+\infty) = 1$ and $F(-\infty) = 0$.

You recall that we have defined limits as $x \to \infty$ or $x \to -\infty$ for a real-valued function of one variable in the Calculus course, MTE–01, Unit 2, Block 1.

Now, you note that $[\omega : X(\omega) < \infty] = \Omega$ and $[\omega : X(\omega) < -\infty] = \phi$ and therefore $P[\omega : X(\omega) < \infty] = P(\Omega) = 1$ and $P[\omega : X(\omega) < -\infty] = 0$.

d) **F(x) is right continuous,**

Recall from your Calculus course (MTE–01, Unit 3) that F(x) is right continuous means that $F(x + h) \to F(x)$ as $x \to 0^+$.

Now, on the basis of these properties can you visualise a distribution function of a random variable graphically ?

Let us first look at some graphs of distribution functions of discrete random variables. Here is an example.

Example 1 : Suppose the random variable X takes the values 0 and 1 with probabilities p and 1 – p, respectively. Then let us obtain the graph of the distribution function of X.

We first note that F(x) is defined for all real x so we must compute $P[X \leq x]$ for both positive and negative real numbers x. Also the smallest value that x can take is 0. Then for any $x < 0$, the event $[X \leq x] = \phi$ for $x < 0$. That is,

$F(x) = P[X \leq x] = 0$ if $x < 0$.

Now consider any real number x greater than or equal to 0 and less than 1. Then the event $[X \leq x]$ for $0 \leq x < 1$ occurs if $X = 1$. That is

$F(x) = P[X \leq x] = P[X = 0] = p$ if $0 \leq x < 1$.

Likewise, if x is a real number greater than or equal to 1, then the event $[X \leq x]$ occurs if $X = 0$ or 1. Therefore

$F(x) = P[X \leq x] = P[X = 0] + P[X = 1] = p + 1 - p = 1$ if $x \geq 1$. Hence the distribution function F(x) is given by

$$F(x) = [X \leq x] = \begin{cases} 0 & \text{if } x < 0 \\ p & \text{if } 0 \leq x < 1. \\ 1 & \text{if } x \leq 1 \end{cases}$$

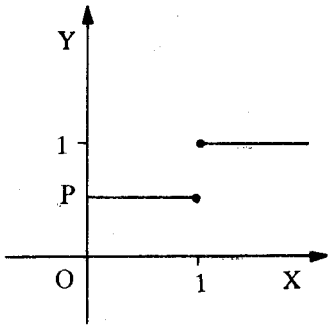


Fig. 1

Now you can easily draw the graph of F(x) as in Fig. 1 (see Fig. 1). What can you say about the continuity of this function ? We leave this as an exercise for you to check (see E1).

E1) What are the points at which the function F(x) given in Example 1 is continuous ?

E2) Suppose X is a random variable taking the values 1, 2 and 3 with probabilities $\frac{1}{6}, \frac{2}{6}$ and $\frac{3}{6}$, respectively. Obtain the distribution function of X and graph it. Also discuss the continuity of the distribution function.

While doing E2 you must have observed the following facts :

i) The graph of F is a step function

ii) The jump discontinuities of F are at the points at which the random variable has positive probabilities.

Now let us look at the graphs of distribution functions in the continuous case.

Example 2 : Suppose the distribution function of a continuous random variable is given by

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \end{cases}$$

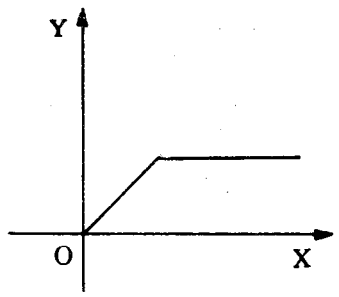


Fig. 2

The graph of F is shown in Fig. 2.

Do you see any difference in the continuity of the functions graphed in Fig. 1 and Fig. 2 ? The graph in Fig. 2 is continuous whereas the graph in Fig. 1 is discontinuous.

Why don't you try some exercises now ?

E3) Graph the following distribution functions and check whether they are continuous or not.

a) $$F(x) = \begin{cases} 0 & , \text{ if } x < 1 \\ 1 - \frac{1}{x^2} & , \text{ if } x \ge 1. \end{cases}$$

b) $$F(x) = \begin{cases} 1 - \frac{1}{3}e^{-2x} & , \ x \ge 0 \\ 0 & , \ x < 0 \end{cases}$$

In E3 (b) you must have seen that the function is neither a pure step function (as in Example 1 and E2), nor a purely continuous function (as in Example 2). The function F has a discontinuity at 0 with a jump of size 2/3 at the point and it is continuous everywhere else.

Now we state the following formulas for computation of probabilities in terms of the distribution function F. The proofs of these formulas are beyond the level of this course.

c) For any x and y,

i) $P[X \leq x] = F(x)$,

ii) $P[X < x] = F(x-0)$,

iii) $P[X = x] = F(x) - F(x-0)$,

iv) $P[x < X \leq y] = F(y) - F(x)$,

v) $P[x \leq X < y] = F(y-0) - F(x-0)$,

vi) $P[x < X < y] = F(y-0) - (F(x)$, and

vii) $P[x \leq X \leq y] = F(y) - F(x-0)$.

If the distribution function F is continuous at a point x, then the limits of F at x from the right and left exist and are both equal to F(x) (see MTE–01, Unit 2). That is, $F(x-0) = F(x+0) = F(x)$, where F(x+0) is the right-hand limit of F at x.

Hence in this case $P[X = x] = F(x) - F(x-0) = 0$.

Thus, if the distribution function F of a random variable X is continuous at a point x, then

$P[X = x] = 0$.

In particular, if the distribution function F is continuous everywhere, then the probability for every singleton, {x} is zero. Inspite of this,

$P[a < X < b] = F(b) - F(a) = P[a < X \leq b] = P[a \leq X < b] = P[a \leq X \leq b]$

for any $a, b \in R$.

Now, suppose the random variable X is discrete and takes the values x_i with $P[X = x_i] = p_i$ for $i \geq 1$. Then from Example 1 and E2 you can see that the distribution function F of X is given by

$$F(x) = \sum_{x_i \leq x} P[X = x_i] = \sum_{x_i \leq x} p_i \ , \quad -\infty < x < \infty$$

where the summation extends over all indices i such that $x_i \leq x$. This distribution function F is a step function (as in Examples 1 and 2). In such a case, F is called a **discrete distribution** and the random variable X is said to be of **discrete type.**

On the other hand, suppose that F is a distribution function of a random variable X whose graph is continuous. For example, the distribution functions in Example 2 and E3 and are continuous. Let us closely look at those graphs. Is there any difference between the graphs ? You might have noticed that graph in E3 is smooth compared to that in Example 2. In mathematical language we say that the distribution function in E3 is not only continuous but it is differentiable.

Note: Henceforth, in this course we will consider only those distribution functions which are differentiable and their derivatives are also continuous (except possibly at discrete set of points, having no effect on any probabilities computed).

That means, there exists a function f defined on the real-line such that

$f(x) = F'(x)$

for all real x. (We shall ignore the points at which the function is not differentiable.) Recall from your Calculus course that such a function F(x) is called an antiderivative

of f(x). Then, since F(x) is continuous by the fundamental theorem of calculus (see Block 3, Unit 1, MTE–01), we have

$$\int_a^x f(t)dt = F(x) - F(a).$$

Taking limits on both sides as $a \to -\infty$, we get,

$$F(x) - F(-\infty) = \int_{-\infty}^{x} f(t)dt.$$

But we have seen that $F(-\infty) = 0$. Therefore we have

$$F(x) = \int_{-\infty}^{x} f(t)dt. \qquad ...(1)$$

Note that f is non-negative since F is non-decreasing.

Summarising our discussion we can say that if F is a distribution function whose derivative exists and is continuous (almost everywhere) on the real-line, then there exists a non-negative function defined on the real line such that

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Such distribution functions are called **absolutely continuous distribution functions** or continuous distributions for short. Therefore, the distribution functions which we shall deal with in this course, are either discrete or (absolutely) continuous. Occasionally, we might consider distribution functions which are neither discrete nor (absolutely) continuous but a mixture of the two as in E7. With some abuse of terminology, **hereafter we shall write continuous distribution for an absolutely continuous distribution.**

You have already studied discrete distribution in Block 2. In the later part of this block, we shall mainly study continuous distributions. The function f(x) which appears in (1) is called density function. In the next section we shall discuss density function in detail.

Before we conclude this section here is an important remark :

Remark 3 : There can be two distinct random variables with the same distribution function. For instance, let us consider the random experiment of tossing an unbiased coin. Define X = 1 if a "head" appears, and X = 0 otherwise. Let Y = 1 if a "tail" appears, and Y = 0 otherwise. Obviously, X and Y are distinct random variables. You can check that both X and Y have the same distribution function.

Now you can check whether you have followed the ideas discussed in this section by attempting the following exercises.

E4) Given the distribution function

$$F(x) = \begin{cases} 0 & \text{for } x < -1 \\ \dfrac{x+2}{4} & \text{for } -1 \le x < 1 \\ 1 & \text{for } x \ge 1, \end{cases}$$

sketch the graph of F and compute

(a) $P\left[-\frac{1}{2} < X \le \frac{1}{2}\right]$ (b) $P[X = 0]$

(c) $P[X = 1]$ (d) $P[2 < X \le 3]$

E5) A random variable X has the distribution function F as shown in the graph given below.

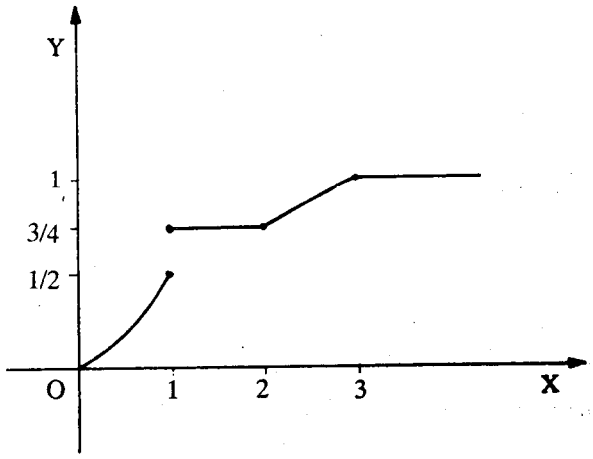


Fig. 3

Find

(a) P[X = 1/2] (b) P[X = 1] (c) P[X < 1]

(d) P[X≤ 1] (e) P[X > 2] (f) P[1/2 < x < 5/2]

Next we shall talk about density functions.

10.3 DENSITY FUNCTIONS

In the last section we said that a distribution function F is absolutely continuous if there is a function f such that

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$

The function f in this expression is called a density function of X. In this section we shall study this density function in detail. We start with its formal definition.

Definition 3 : A function f defined on the real-line is called a **density function** of a random variable X if

(i) $f(x) \geq 0$ for all x

(ii) $P[a < X \leq b] = \int_{a}^{b} f(y)\, dy$, for all a b, $\in \mathbf{R}$ and $a \leq b$,

In particular, observe that $\int_{-\infty}^{\infty} f(y)\, dy = 1$.

Now, suppose that F is a distribution function such that F′ exists and F′ is continuous. Then we know that

$$F(x) = \int_{-\infty}^{x} f(y)\, dy.$$

for some non-negative real-valued function f. Now let us verify whether f satisfies (i) and (ii) in Definition 3. (i) is automatically satisfied. To verify (ii), note that

$$P[a < X \leq b] = F(b) - F(a)$$

$$= \int_{-\infty}^{b} f(y)dy - \int_{-\infty}^{a} f(y)\, dy$$

$$= \int_a^b f(y)\, dy.$$

Therefore, f satisfies the conditions (i) and (ii).

Conversely, if f is a density function of a random variable X, then define

$$F(x) = \int_{-\infty}^{x} f(y)\, dy.$$

Then, by (ii) $F(x) = P[X \leq x]$. Therefore, F is a distribution function of the r.v. X.

Also, by using the Fundamental Theorem of Calculus (Theorem 7, Unit 10, MTE-01), F is differentiable and $\frac{dF}{dx} = f(x)$. Further

$$P[a < X \leq b] = \int_a^b f(y)\, dy$$

for any pair of real numbers a and b.

Again, from Unit 15, MTE-01, you know that the integral of a function f between the limits a and b can be interpreted as the area bounded between the curves $y = f(x)$, the x-axis and the ordinates $y = a$ and $y = b$. Hence this area is equal to the probability that the random variable X takes values between a and b.

Note that the area enclosed between the curve $y = f(x)$ and the line $y = 0$ is unity, since it is equal to $P[-\infty < X < \infty]$.

Let us now look at some examples of density functions and their corresponding distribution functions.

Example 3 : Let X be a random variable with density function

$$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Then

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(y)\, dy$$

i.e.,

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ x & \text{for } 0 < x < 1 \\ 1 & \text{for } x \geq 1 \end{cases}$$

You can see the graphs of f and F in Fig. 4

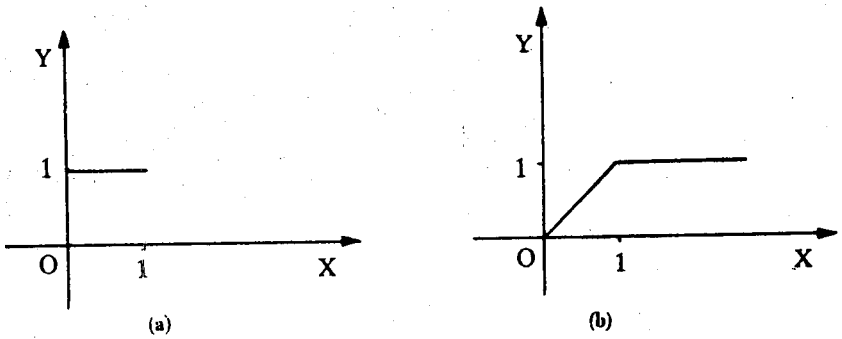


Fig. 4 : Graph of (a) density function and (b) corresponding distribution function

This distribution is called the standard **uniform distribution** or rectangular **distribution.**

Note that, for any a and b with $0 \le a < b \le 1$,

$$P[a \le X \le b] = \int_a^b dy = b - a,$$

Hence the probability that a real number is selected from [a, b] under this probability model is b – a. It is just the length of the interval [a, b].

Let us consider another example.

Example 4 : Let X be a random variable with density function f which is a constant over an interval $[\alpha, \beta]$ and equal to zero outside the interval $[\alpha, \beta]$. In other words

$$f(x) = \begin{cases} C & \text{for } \alpha \le x \le \beta \\ 0 \ , & \text{otherwise,} \end{cases}$$

where C is a constant. From the properties of density function, we get $C \ge 0$. Further the equation

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

implies that

$$\int_{\alpha}^{\beta} C\, dx = 1.$$

This relation leads to $C = \frac{1}{\beta - \alpha}$ and we have

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 0 \ , & \text{otherwise} \end{cases}$$

This is called the **uniform density function on the interval** $[\alpha, \beta]$. The corresponding distribution function is

$$F(x) = \begin{cases} 0 & \text{if } x < \alpha \\ \frac{x - \alpha}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 1 & \text{if } x > \beta \end{cases}$$

called the **uniform distribution on** $[\alpha, \beta]$. Did you notice that Example 3 is a particular case of Example 4? We trust you will be able to check the calculation of F from f very easily.

In the next example we discuss another distribution which is frequently used as a model in describing the life time of a light bulb.

Example 5 : Suppose X denotes the life time of a bulb and X has density function

$$f(x) = \begin{cases} e^{-x} & , x \ge 0 \\ 0 & , x < 0. \end{cases}$$

Check that $f(x) \ge 0$ and that $\int_{-\infty}^{\infty} f(x)\, dx = 1$. We now claim that the distribution function

F corresponding to f is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-x} & \text{for } x \geq 0. \end{cases}$$

Do you agree? Check that $\frac{dF(x)}{dx} = f(x)$, and you will be convinced. This distribution is known as the **standard exponential distribution.**

Another distribution which is by far the single most important distribution in Statistics is the **normal distribution.** It is sometimes referred to as the Gaussian distribution. We take this up in our next example.

Example 6 : Suppose X is a random variable with density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

It is obvious that f is a non-negative function. It needs some effort to show that

$$\int_{-\infty}^{\infty} \phi(x)dx = 1.$$

We will postpone this proof until Unit 11. The distribution function F corresponding to this density function is

φ and Φ are the greek letters small phi and capital phi.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy, \quad -\infty < x < \infty.$$

We have sketched the graphs of φ and Φ in Fig. 5 below.

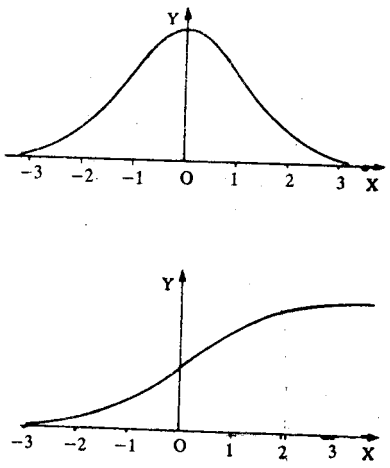


Fig. 5 : Graph of the (a) density function, b) distribution function of a standard normal distribution

We will study this graph in detail in the next unit.

The distribution discussed in the above example is known as the **standard normal distribution.**

See if you can solve these exercises now.

E6) Suppose that a random variable X has the density function

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < \infty$$

Find the value x_0 such that $F(x_0) = .5$.

E7) A random variable X has the distribution function

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ x^2 & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x \ge 1. \end{cases}$$

Show that X is of continuous type, and determine its density function.

E8) Show that the function

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty$$

is a density function.

E9) Buses arrive at a specified stop at 15-minute intervals starting at 8.00 A.M. That is, they arrive at 8.00 A.M., 8.15 A.M., 8.30 A.M., and so on. If a passenger arrives at the stop at a time that is uniformly distributed between 8.00 A.M. and 8.30 A.M., find the probability that she waits less than 5 minutes for a bus.

E10) Consider the function

$$f(x) = C(2x - x^3), \quad 0 < x < \frac{5}{2}$$

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

Can f be a probability density function? If so, calculate the constant C.

By now you must have become quite familiar with the density and distribution functions of a random variable. In the next section we take up the study of the expectation, variance and other related concepts for a r.v.

10.4 EXPECTATION AND VARIANCE

In Block 1 you have calculated the mean (expected value), variance and other moments of a frequency distribution of a quantitative character. In Block 2, again, you have studied these very concepts in the context of discrete probability distributions. If you remember, over there you had replaced relative frequencies by probabilities. Now we are going to study these concepts again — this time for a continuous random variable. Since you are already familiar with the interpretations and interrelationships of these concepts, here we shall go over them quickly. Quite often we'll only state the all-too-familiar results and expect you to prove them. Let us start with the definition.

Definition 4 : The **expectation** of a r.v. X with density function f is defined to be

$$\int_{-\infty}^{\infty} x\, f(x)\, dx.$$

provided $\int_{-\infty}^{\infty} |x|\, f(x)\, dx < \infty.$

We denote the expectation or expected value of X by E(X) whenever it exists.

In general, if g is a function of the r.v. X, we define

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx.$$

We will discuss a function of a r.v. in more detail in Sec. 10.6.

provided $\int_{-\infty}^{\infty} |g(x)|\, f(x)\, dx < \infty.$

With this general definition we'll be able to write down the expression for the variance and the moments of a r.v. X.

As you know, from Block 1,

$$\text{Var}(X) = E\left[(X - \mu)^2\right], \text{ where } \mu = E(X).$$

Therefore, we write

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx,$$

provided the integral on the R.H.S. is finite.

Then using some algebraic properties of expectation we can show that $\text{Var}(X) = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2$. We shall discuss this at the end of this section. Variance is also denoted by σ^2 and we are sure you remember that σ is called the **standard deviation.** Have you noted the similarities and the dissimilarities between these definitions and those given in Blocks 1 and 2? A major point of dissimilarity is that here we have defined the expected values as integrals, whereas earlier we had used summations. But hadn't you expected this ? Since our random variable now varies continuously, instead of taking only discrete values, it is quite natural that we use integrals and not summations. Another change is that the density function f(x) now takes the place of the p.m.f. But these differences apart, don't you agree that the basic concept remains the same ?

Before we talk about the algebraic properties of expectation and variance, we give a few examples. These will familiarise you with the calculations of mean and variance.

Example 7 : Let us calculate the expected value and the variance of X, where X is a r.v. with uniform distribution on $[\alpha, \beta]$ described in Example 6.

Now,

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & , \quad \text{if } \alpha \le x \le \beta \\ 0 & , \quad \text{otherwise} \end{cases}$$

By definition,

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\ &= \int_{\alpha}^{\beta} x \frac{1}{(\beta - \alpha)}\, dx, \text{ since } f(x) = 0 \text{ outside } [\alpha, \beta] \\ &= \frac{1}{\beta - \alpha}\left[\frac{\beta^2 - \alpha^2}{2}\right] \\ &= \frac{\alpha + \beta}{2}. \end{aligned}$$

Thus, the expected value of X is the mid-point of the interval $[\alpha, \beta]$.

If X is a discrete r.v., then $E(X) = \Sigma\, xp(x)$,

Now, $\text{Var}(X) = E(X^2) - \left[\dfrac{\alpha + \beta}{2}\right]^2$,

and

$$\begin{aligned} E(X^2) &= \int_{\alpha}^{\beta} x^2 \frac{1}{\beta - \alpha}\, dx \\ &= \frac{1}{\beta - \alpha}\left[\frac{\beta^3 - \alpha^3}{3}\right] \end{aligned}$$

Hence, $\text{Var}(X) = \dfrac{\beta^3 - \alpha^3}{3(\beta - \alpha)} - \dfrac{(\alpha + \beta)^2}{4}$

$$= \frac{\beta^2 + \alpha\beta + \alpha^2}{3} - \frac{(\alpha + \beta)^2}{4}$$

$$= \frac{(\beta - \alpha)^2}{12}.$$

Let us consider another example.

Example 8 : Suppose X is a r.v. with exponential distribution. This means that the density function f of X is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad x \geq 0 \\ 0 & , \quad x < 0, \end{cases}$$

where λ is some positive constant.

Recall that we have seen the case $\lambda = 1$ in Example 5.

Let's compute the mean and variance of X.

$$E(X) = \int_0^\infty x \lambda e^{-\lambda x} dx$$

$$= \frac{1}{\lambda} \int_0^\infty y e^{-y} dy \quad , \text{ if we put } y = \lambda x.$$

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

Do you agree that $\int_0^\infty y e^{-y} dy = 1$? Note that this follows by the method of integration by parts. If X is interpreted as the life-time of an electric bulb (see Example 5), then the mean or expected life time is $1/\lambda$. Now, to calculate Var(X), we begin by computing $E(X^2)$.

$$E(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x} dx$$

$$= \frac{1}{\lambda^2} \int_0^\infty y^2 e^{-y} dy, \text{ where } y = \lambda x.$$

You may not have come across this integral before. It is the value of the **gamma function** at 3. The gamma function, Γ, is defined as

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha - 1} e^{-y} dy, \text{ where } \alpha > 0.$$

Then it is known that

$\Gamma(n + 1) = n !$, where n is any non-negative integer. Without going into the how and why of this, we shall only use this fact to evaluate $E(X^2)$.

So, $$E(X^2) = \frac{1}{\lambda^2} \Gamma(3).$$

$$= \frac{1}{\lambda^2} 2 !$$

$$= \frac{2}{\lambda^2}$$

Thus, $\quad \mathrm{Var}(X) = E(X^2) - [E(X)]^2$

$$= \frac{1}{\lambda^2}.$$

In all the examples considered so far, the random variables turned out to have finite expectations. But you should note that there are cases of r.v.s. whose expectations not exist. You can see one such r.v. in the next example.

Example 9 : Let X be a r.v. with density function,

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

In E8 you must have proved that the function f above is a density function. **This** distribution is called the (standard) **Cauchy distribution.**

Let us check whether E(X) exists in this case.

We have

$$\int_{-\infty}^{\infty} |x|\, f(x)\, dx = 2\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\, dx$$

$$= \frac{1}{\pi}[\ln(1+x^2)]_0^{\infty}$$

and $\frac{1}{\pi}\ln(1+x^2) \to \infty$ as $x \to \infty$. Hence, $\int_{-\infty}^{\infty} |x|\, f(x)\, dx$ is not finite. **Therefore, E(X)** does not exist.

From the above examples you must have got a pretty good idea of the computations required to evaluate the expectation and the variance of a r.v. Now we shall list some of their algebraic properties. The proofs of these properties depend on some elementary properties of integrals. We have proved some of them and we are sure you will be able to prove the rest (see E11).

i) If $Y = aX + b$, where a and b are any two constants, and if X has a finite expectation then E(Y) exists and
$E(Y) = a\,E(X) + b.$

ii) If X is a r.v. taking non-negative values with probability one, and if $E(X) < \infty$, then $E(X) \geq 0$.

iii) If X and Y are r.v.'s with finite expectations and a and b are constants, then
$E(aX + bY) = a\,E(X) + b\,E(Y)$
We will come back to the proof of this property in Unit 12.
But note that this result can be extended to three or more variables.

iv) $\mathrm{Var}(X) = 0$ if and only if X is a constant with probability one, i.e. iff $P[X = C] = 1$ for some constant C.

One-way implication in this statement is easy. Then we can consider X as a discrete r.v. So let's apply the definition of the expected value of a discrete r.v. to get E(X). You will see that we get $E(X) = C$. Then $\mathrm{Var}(X) = E[X - C]^2 = 0$, since $(X - C)^2 = 0$ with probability one.

To prove the converse, we need to use Chebyshev's inequality. So, we postpone the proof till Unit 14, where we are going to discuss this inequality.

v) For any two constants a and b,

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$$

vi) Var (X) = $E(X^2) - \mu^2$, where $\mu = E(X)$.

We have already mentioned this property earlier.

Now attempt the following exercises to complete the discussion of algebraic properties of E(X) and Var (X).

E11) Prove the properties i), v) and vi) above.

E12) Compute the expectation and variance of a random variable Y whose density function is

$$f(y) = \begin{cases} 1 - |y| & \text{for } 1 - |y| < 1 \\ 0 & \text{otherwise} \end{cases}$$

E13) Let X be a random variable with density function

$$f(x) = \begin{cases} \frac{2}{x^3} & \text{if } x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Show that E(X) exists and E(X) = 2 but Var(X) does not exist.

E14) Let X be a random variable such that $E\left[(X - a)^2\right]$ exists for all real numbers a. Show that $E\left[(X - a)^2\right]$ is minimum when $a = \mu = E(X)$.

So far we have seen how to calculate the expectation and variance of a random variable as long as they exist. As we have seen, both the expectation and variance are specified by the values E(X) and $E(X^2)$, the expected values of X raised to the first and second powers. But these two expected values describe only two particular aspects : "the middle value" and the measure of relative variability about the middle value of the probability distribution corresponding to the random variable. These two numbers are not sufficient to describe the distribution completely. To get more information about the distribution we need to study its moments which are specified by the values of $E(X^k)$, k = 1, 2, 3,..... We shall take up this in the next section.

10.5 MOMENTS AND MOMENT GENERATING FUNCTION

In Unit 7, we have discussed moments and m.g.f. of a discrete r.v. X. The discussion in the case of a continuous probability distribution runs parallel to that in the case of discrete probability distributions.

If k is any integer greater than or equal to one, and if b is any real number, then if $E[X - b]^k$ exists, it is called the **k^{th} moment of X about the point b.**

kth moments about b = 0 given by

$$\mu_k' = E(X^k), \quad k = 1,2,...$$

are called **raw moments** or simply **moments.** Now if we take $b = \mu = E(X)$, then

$$\mu_k = E\left[(X - \mu)^k\right], \quad k = 1,2,...$$

which are the moments about the mean μ, are called **central moments.**

k is called the **order of the moment** $(X - b)^k$.

Do you agree with the following observations ?

$$\mu_1' = \mu$$
$$\mu_1 = 0$$
$$\mu_2 = \sigma^2 = \text{Var}(X)$$

In Blocks 1 and 2, we had derived the relations between raw and central moments of X. The same hold good here. Thus, we have

$$\mu_k = \sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} \mu_1' (\mu_1')^{k-i}, \mu_0' = 1.$$

At the end of the last section we said that we need to study moments to get more information about the distribution of a r.v. You may think, what additional information can the moments give ?

To see that let us look at the following expressions.

$$\gamma_2 = \frac{E|(X-\mu)^3|}{\sigma^3} = \frac{\mu_3}{\sigma^3} \text{ and } \beta_2 = \frac{E|(X-\mu)^4|}{\sigma^4} = \frac{\mu_4}{\sigma^4}.$$

Now what is the significance of this? Aren't the expression for γ_1 and β_2 familiar ? γ_1 measures the skewness and β_2, the kurtosis of a density function. (Compare these with the measures of skewness and kurtosis of a frequency distribution, discussed in Unit 3.)

Let us see an example.

Example 11 : Suppose X has uniform distribution on $[\alpha . \beta]$. Let us compute the raw moments for this distribution.

We have,

$$E(X^r) = \int_{\alpha}^{\beta} x^r \frac{1}{\beta-\alpha} dx$$

$$= \frac{1}{\beta-\alpha}\left[\frac{x^{r+1}}{r+1}\right]_{\alpha}^{\beta} = \frac{\beta^{r+1}-\alpha^{r+1}}{(r+1)(\beta-\alpha)}$$

In particular

$$\mu = E(X) = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{\beta+\alpha}{2}.$$

Why don't you try some exercises now ?

E15) Suppose X has uniform distribution on $[\alpha, \beta]$. Find $E(X^r)$ and $E[(X-\mu)]^r$ for $r \geq 1$, where μ is the mean of X.

E16) Suppose X has a standard exponential density. Find the coefficient of skewness.

E17) If Y = ax+b, show that Y has the same coefficients of skewness and kurtosis as X, whenever they exist.

After doing these exercises you would have realised that calculation of moments of a random variable is cumbersome even when they exist. Alternatively, we can use the moment generating function, whenever it exists, to obtain the moments.

Let X be a random variable, such that $M_X(t) = E[e^{tX}]$ exists for some $t \neq 0$. $M_X(t)$ is called the **moment generating function** (m.g.f) of the random variable X, whenever it is well-defined.

Note that $M_X(0) = 1$ for any random variable X. Let us expand $M_X(t)$ by Maclaurin's series expansion (See Unit 6, Block 2 of MTE-01). Then we have

$$M_X(t) = M_X(0) + t\frac{dM_X(t)}{d}\Big|_{t=0} + \ldots\ldots + \frac{t^n}{n!}\frac{d^n}{dt^n}(M_X(t))\Big|_{t=0} + \ldots\ldots \quad \ldots\ldots(2)$$

On the other hand, suppose the following computation is justified :

$$E[e^{tX}] = E\left[1 + tX + \frac{t^2X^2}{2!} + \ldots\ldots + \frac{t^nX^n}{n!} + \ldots\ldots\right]$$

$$= 1 + t\,E(X) + \frac{t^2}{2!}E(X^2) + \ldots\ldots + \frac{t^n}{n!}E(X^n) + \ldots\ldots$$

.....(3)

Comparing the coefficient of t^n for every n = 1, 2, 3 ...in (2) and (3), we have the relation

$$\left.\frac{d^n}{dt^n}(M_X(t))\right|_{t=0} = E(x^n),\ n \geq 1.$$

This relation implies that the nth moment about zero of the random variable X can be obtained by differentiating the m.g.f. $M_X(t)$ of X exactly n times, and then evaluating the nth derivative at zero. This is why $M_X(t)$ is called a "moment generating function" of X. We can justify the above arguments under some conditions on the existence of moments of X. But this discussion is beyond the scope of this course.

We now show you how to calculate the m.g.f. for the uniform, and the exponential distributions.

Example 11 : Suppose X has a uniform distribution on $[\alpha, \beta]$. Let us compute the m.g.f. of this distribution.

We have

$$M_X(t) = E[e^{tX}] = \int_\alpha^\beta \frac{e^{tx}}{\beta - \alpha}\,dx$$

$$= \frac{1}{\beta - \alpha}\left[\frac{e^{tx}}{t}\right]_\alpha^\beta \quad \text{for } t \neq 0$$

$$= \frac{e^{t\beta} - e^{t\alpha}}{t(\beta - \alpha)} \quad \text{for } t \neq 0$$

and

$$M_X(0) = 1.$$

Hence the m.g.f. $M_X(t)$ exists for all t.

Now, why don't you check your answers to E15 by calculating the moments from the m.g.f. obtained in this example ? (see E18).

Example 12 : Suppose X has the exponential density,

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & , \quad x > 0 \\ 0 & , \quad x \leq 0 \end{cases}$$

where $\lambda > 0$. Let us compute the m.g.f.

$$M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx}\,\lambda\,e^{-\lambda x}\,dx$$

$$= \lambda\int_0^\infty e^{(t-\lambda)x}\,dx$$

$$= \frac{\lambda}{\lambda - t} \qquad \text{for } t < \lambda.$$

This m.g.f. $M_X(t)$ does not exist for $t \geq \lambda$ since in that case, $e^{(t-\lambda)x}$ is unbounded on $]0, \infty[$.

Try to solve these exercises now.

E18) Calculate the first and second moments of the uniform distribution using the m.g.f. of the distribution.

E19) Let X be a random variable with density function

$$f(x) = \begin{cases} x & \text{if } 0 \le x \le 1 \\ 2-x & \text{if } 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Determine the m.g.f. $M_X(t)$ of X whenever it exists.

E20) Suppose X have the density function

$$f(x) = \begin{cases} xe^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find its moment generating function whenever it exists.

E21) Let $Y = aX + b$. Show that

$M_Y(t) = e^{bt} M_X(at)$.

Here we make an important remark.

Remark 4 : You may have got the wrong impression that if two r.v.s X and Y are such that all the moments of X are respectively equal to the moments of Y, then X and Y are equal. This is not so. Infact moments do not even determine a distribution uniquely. Same is the case in general with m.g.f.s. However the following theorem is valid. We will not discuss its proof as it is beyond the scope of this course.

Theorem 1 : Suppose X and Y are random variables with m.g.f.s $M_X(t)$ and $M_Y(t)$, respectively. Suppose $M_X(t)$ and $M_Y(t)$ exist in an open interval containing zero and $M_X(t) = M_Y(t)$ in that interval. Then X and Y have the same distribution.

In the next section, we take up one last topic, the distribution of a function of a random variable.

10.6 FUNCTIONS OF A RANDOM VARIABLE

In the earlier sections of the units we were concerned with various aspects of the distribution function of a random variable. In many applications we may have to consider not only the distribution function of a random variable, but the distribution function of a function of a random variable. In this section we first try to understand what is meant by a function of a random variable and then discuss how to find its distribution function.

Let us first consider this situation :

Suppose we want to know the volume $V = \frac{4}{3}\pi r^3$ of a spherical object, say ball bearing, manufactured by a company. Due to manufacturing defect, the radii of different spheres may be different. We suppose that the radius of a sphere is a continuous random variable X having density function f. Then we can consider V as function of the random variable X, say

$$V = \frac{4}{3}\pi X^3$$

Here we would expect that we can derive the density function of V from the knowledge of the density function of X. In such situations we are concerned with the

concept of a function of a random variable and its density function. Formally we define the concept as follows :

Definition 5 : Let X be an r.v. defined on Ω and $g : R \to R$. Then the real valued function Y defined on Ω by

$$Y(\omega) = g[X(\omega)]$$

is called a **function of the random variable X.**

For example, if X is an r.v. and $g(x) = ax + b$, then $Y = aX + b$ is a function of the r.v. X.

You have already come across some functions of r.v.s in the earlier sections like X^2, X^3, ... In this unit we consider only the continuous case. Here we make some remarks.

Remark 5 : a) In general a function of a r.v. need not be an r.v. But it turns out that whenever g has nice properties, some of which are continuity (and) monotonocity, then Y becomes an r.v. So, in this course, whenever we deal with functions of r.v., we assume that g has nice properties by which Y becomes an r.v.

b) Another question which comes to our mind is that suppose X is a continuous (discrete) r.v., is it true that Y is also continuous (discrete) ? In general we cannot conclude from the definition that Y is of the same type as X. For instance, if $g(x) = c$, a constant, for all $x \in R$, then $P[Y = c] = 1$ and Y is a degenerate r.v. (Recall the definition of a degenerate r.v. from Unit 7, Block 2.)

Next we shall see how we find the distribution of $Y = g(X)$. Let us first consider a simple case :

Suppose X is a continuous r.v. and $g(x) = ax + b$, where $a, b \in R$, $a > 0$. Then $Y = aX + b$.

To get the distribution of Y, we consider

$$\begin{aligned} P[Y \le y] &= P[aX + b \le y] \\ &= P[X \le a^{-1}(y - b)] \end{aligned}$$

Now if F_Y and F_X denote the distribution functions of Y, and X respectively, then we have

$$F_Y(y) = F_X\left[a^{-1}(y - b)\right]$$

Differentiating both sides with respect to y, we get the density of Y as

$$f_Y(y) = f_X\left[a^{-1}(y - b)\right]$$

In this case we could easily derive the density function because the inverse of g exists (given by $g^{-1}(y) = a^{-1}(y - b)$), and the inverse is differentiable. So, in the general case we expect that if g has similar properties as in the above case, then we can find the density function.

Now let us start with a real-valued function g defined on R. Let $Y = g(X)$.

We call $D = \{x : f(x) > 0\}$ as the **support** of f.

Let us now suppose that g is a continuous and strictly increasing function on the support of f. Since g is strictly increasing, there exists a function s, such that

$$g[s(y)] = y$$

for all y . s is called the **inverse** function of g. Since g is continuous, s is also continuous.

Further $g(x) \le y$ if and only if $x \le s(y)$. Hence

$$P[Y \le y] = P[g(X) \le y]$$
$$= P[X \le s(y)]$$

Therefore, if F_X and F_Y denote the distribution functions of X and Y respectively, then

$$F_Y(y) = F_X(s(y))$$

In particular, suppose X has a density function f_X and s(y) is differentiable in y. The

Y = g(X) has a density function, f_Y and

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(s(y))}{dy}$$
$$= f_X(s(y)) \left[\frac{d}{dy}[s(y)]\right]$$

Now on the other hand if g is continuous and strictly decreasing on the support of f, then also we can use a similar argument as in the earlier case. To see this, first note that g has a unique inverse function s(y) which is continuous and strictly decreasing. Hence $g(x) \le y$ if $x \ge s(y)$. Therefore,

$$F_Y(y) = P[Y \le y] = P[g(x) \le y]$$
$$= P[X \ge s(y)]$$
$$= 1 - P[X < s(y)]$$
$$= 1 - F_X[s(y) - 0]$$

If X has a density function f, then F_Y is continuous and

$$F_Y(y) = 1 - F_X(s(y))$$

In fact, Y has a density function $f_Y(y)$ given by

$$f_Y(y) = \frac{dF_Y(y)}{dy} = -\frac{dF_X(s(y))}{dy}$$
$$= -f_X(s(y)) \frac{d}{dy}[s(y)]$$

Thus we have proved the following theorem.

Theorem 2 : Suppose X is a random variable with density function f_X. Let Y = g(X), where g is continuous, and either strictly increasing or strictly decreasing function. Let x = s(y) be the inverse function of g and suppose s is differentiable in y. Then Y has a density function $f_Y(y)$ and

$$f_Y(y) = f_X[s(y)] \left|\frac{d}{dy}[s(y)]\right|$$

Let us consider some examples.

Example 13 : Suppose X has a density function

$$f(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then the support of f is]0, 1[.

Suppose $Y = X^2$. Here $g(x) = x^2$ in the earlier notation.

Note that g is strictly increasing and continuous on the]0, 1[and $s(y) = y^{1/2}$ is the inverse of g on]0, 1[. Then s(y) is differentiable in]0, 1[and therefore by Theorem 2, we have

$$f_Y(y) = f[s(y)] \frac{d}{dy}[s(y)] = \begin{cases} 3[y^{1/2}]^2 \frac{1}{2} y^{-1/2}, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Thus, the density function of Y is

$$f_Y(y) = \begin{cases} \frac{3}{2} y^{1/2} & , \quad 0 < y < 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Example 14 : Suppose X is a random variable with standard uniform density function, that is

$$f_X(x) = 1, \ 0 < x < 1$$
$$= 0, \ \text{otherwise}$$

Define $Y = \frac{1}{\lambda} \ln X$. The function

$$g(x) = \frac{-1}{\lambda} \ln x$$

maps the interval]0, 1[to]0, ∞[. Further g(x) is continuous and strictly decreasing on the interval]0, 1[. The inverse function s(y) of g is given by

$$s(y) = e^{-\lambda y}, \quad 0 < y < \infty$$

which is differentiable. Therefore by Theorem 2, Y has a density function given by

$$f_Y(y) = -f[s(y)] \frac{d}{dy}[s(y)]$$

$$= \begin{cases} \lambda e^{-\lambda y} & \text{for } y > 0 \\ 0 & \text{for } y \le 0. \end{cases}$$

Sometimes we cannot apply Theorem 2 which we have used in the two examples above. The next example gives one such situation.

Example 15 : Suppose X has the standard normal density function. Let $Y = X^2$. Here $g(x) = x^2$. This function is continuous but strictly increasing on [0, ∞[and strictly decreasing on]−∞, 0[. Further g(x) is not one-to-one, since g(−x) = g(x). That means g(x) does not have an inverse. Therefore we cannot apply Theorem 2 in this case. So we try some other method in this case. Since $y \ge 0$ for all x, we have

$P[Y \le y] = P[X^2 \le y] = 0$ for $y < 0$ and for $y > 0$,

$P[Y \le y] = P[X^2 \le y] = P[|X| \le \sqrt{y}] = P[-\sqrt{y} \le X \le \sqrt{y}]$

i.e. $$F_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} f(x)\, dx,$$

where f is the standard normal density function. By the symmetry of this density function, for y > 0 we have

$$F_Y(y) = 2 \int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

$$= 2 \int_0^{y} \frac{1}{\sqrt{2\pi}} e^{-z/2} \frac{dz}{2z^{1/2}}$$

$$= \int_0^{y} \frac{1}{\sqrt{2\pi}} z^{-1/2} e^{-z/2}\, dz.$$

This derivation proves that the random variable Y has a density function f_Y, where

$$f_Y(y) = \begin{cases} 0 & \text{for } y \le 0 \\ \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} & \text{for } y > 0. \end{cases}$$

Recalling that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, we can write $f_Y(y)$ in the form,

$$f_Y(y) = \begin{cases} 0 & \text{for } y \le 0 \\ \frac{1}{2^1/2\,\Gamma(1/2)} y^{(1/2)-1} e^{-y/2} & \text{for } y > 0. \end{cases}$$

This density function is known as a **Chi-square** density with 1 degree of freedom. We will study more about this distribution in Unit 13.

Now, suppose we want to compute the expectation of $Y = g(X)$ whenever it exists. We can either use the distribution of X or the distribution of Y. For instance, suppose X has a density f_X and Y has density f_Y.

Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

and

$$E(Y) = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy .$$

It can be shown that E(Y) exists if and only E[g(x)] exists and both these methods of calculation lead to the same result. We do not give the reasoning here. The choice of the method depends on the complexity involved in finding the distribution or the density of $Y = g(X)$.

Let us continue the discussion in Example 15 to illustrate this.

Example 16 : In Example 15, you have seen that $E(X^2) = 1$. Let us compute E(Y) where $Y = X^2$ directly using the probability density of Y derived above.

Then

$$E(Y) = \int_0^{\infty} y \frac{1}{2^{1/2}\,\Gamma(1/2)} y^{(1/2)-1} e^{-y/2}\, dy$$

$$= \frac{1}{2^{1/2}\,\Gamma(1/2)} \int_0^{\infty} y^{3/2-1} e^{-y/2}\, dy$$

$$= \frac{1}{2^{1/2}\,\Gamma(1/2)} 2^{1/2} \int_0^{\infty} 2\, u^{3/2-1} e^{-u}\, du$$

$$= \frac{2}{\Gamma(1/2)} \Gamma(3/2) .$$

But $\Gamma(3/2) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right)$. Hence E(Y) = 1 as it should be.

Now it's time to do some exercises.

E22) Find the density function of $Y = X^2$ when X has a uniform density on [–1, 1].

E23) Suppose a random variable X has the density function

$$f(x) = \begin{cases} \frac{x}{2} & 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = 4 - x^3$. Find the density function of Y.

We now end this discussion. We hope that by now you would have gained reasonable knowledge about the various aspects related to the distribution of a random variable. In the next unit we shall study some standard distributions.

10.7 SUMMARY

In this unit, we have

1) introduced the concepts of the distribution function of a random variable and the density function for an (absolutely) continuous distribution;
2) studied properties of a distribution function and a density function;
3) defined the notions of moments of a random variable in general and the expectation (mean) and the variance in particular ;
4) introduced the concept of a moment generating function for a random variable; and
5) given methods for finding the distribution function or the density function of a function of a random variable.

You may now like to go back to Sec. 10.1 and go through the list of **unit objectives** to see if you have achieved them. If you want to see what our solutions to the exercises in the unit are, we have given them in the following section.

10.8 SOLUTIONS AND ANSWERS

E1) a) 0 and 1 are the points at which the function is discontinuous.

b) At 0, the function has a jump discontinuity of size p and at 1, the function has a jump discontinuity of size 1– p.

E 2) The distribution function $F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/6 & \text{if } 1 \le x < 2 \\ 3/6 & \text{if } 2 \le x < 3 \\ 1 & \text{if } x \ge 3 \end{cases}$

The graph of F is as in Fig. 6

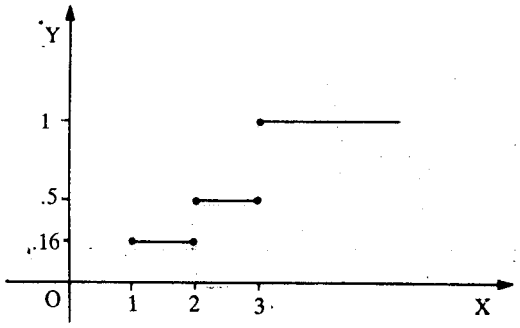


Fig. 6.

The function has discontinuity at x = 1, at x = 2 and at x = 3.

E3) a) Fig. 7 shows the graph of F(x).

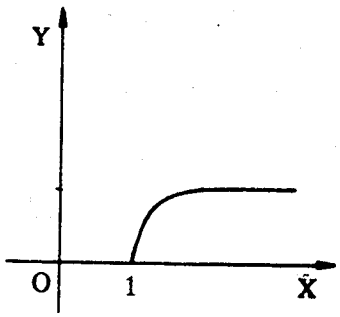


Fig. 7

The graph shows that F(x) is continuous for all x.

b) Fig. 8 shows the graph of F(x).

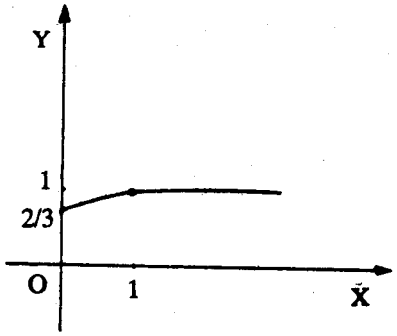


Fig. 8

The graph shows that the function is discontinuous at x = 0 and is continuous everywhere else.

E4)

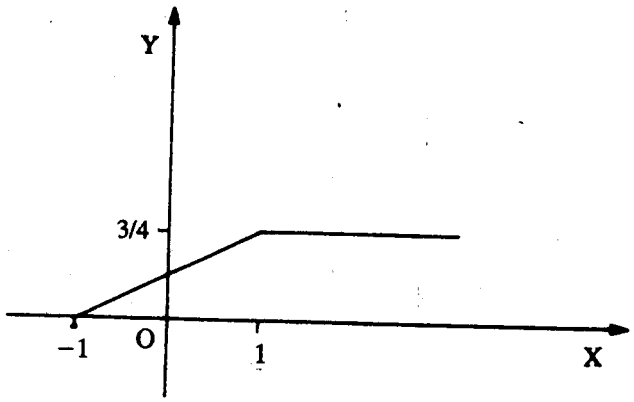


Fig. 9

a) By property (e), we have

$$P\left[-\frac{1}{2} < X \le \frac{1}{2}\right] = F\left(\frac{1}{2}\right) - F\left(\frac{1}{2} - 0\right)$$
$$= \frac{5}{8} - \frac{3}{8} = \frac{1}{4}$$

b) 0

c) $\frac{1}{4}$

d) 0.

E5) From the figure we get that

a) $F(x - 0) = 0$, if $x < 0$
$= \frac{x^2}{2}$, if $0 \le x < 1$
$= \frac{3}{4}$, if $1 \le x < 2$
$= x$, if $2 \le x < 3$
$= 1$, if $x \ge 3$.

Then by Property (e) we have

$$P\left[X = \frac{1}{2}\right] = F\left(\frac{1}{2}\right) - F\left(\frac{1^-}{2}\right)$$
$$= F\left(\frac{1}{2}\right) - F\left(\frac{1}{2}\right), \quad \text{since the function is continuous at } x = \frac{1}{2}$$
$$= 0$$

b) $P[X = 1] = F(1) = F(1 - 0)$
$= \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$.

Similarly we get

c) $P[X < 1] = \frac{1}{2}$

d) $P[X \le 1] = \frac{3}{4}$

e) $P[X > 2] = \frac{1}{4}$

f) $P\left[\frac{1}{2} < X < \frac{5}{2}\right] = \frac{3}{4}$

E6) The density function of X is

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < +\infty$$

Then the distribution function of F (x) is given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{2} e^{-|x|} dx$$
$$= \frac{1}{2}\left[\int_{-\infty}^{0} e^{x} dx + \int_{0}^{x} e^{-x} dx\right]$$
$$= \frac{1}{2}\left[\left\{e^{x}\right\}_{-\infty}^{0} + \left\{e^{-x}\right\}_{0}^{x}\right]$$
$$= \frac{1}{2}[1 + e^{-x} - 1]$$

$$= \frac{1}{2} e^{-x}$$

$$= .5e^{-x}$$

Then the point x_0 satisfying $F(x_0) = .5$ is $x_0 = 0$.

E7) The graph of F(x) is shown in Fig. 10.

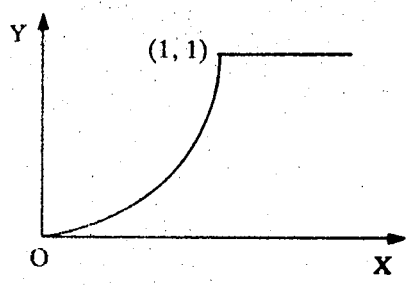


Fig. 10

The graph shows that the function F(x) is continuous for all x.

The density function f(x) is given by

$$f(x) = \frac{dF(x)}{dx} = 2x \quad , \text{ if } 0 < x < 1$$

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

E8) The given function is $f(x) = \frac{1}{\pi(1 + x^2)}$, $-\infty < x < \infty$.

Then f(x) satisfies the following conditions.

i) $f(x) \geq 0$ for all x.

ii) $$\int_{-\infty}^{+\infty} \frac{1}{\pi(1 + x^2)} = 2\int_{0}^{+\infty} \frac{1}{\pi(1 + x^2)} dx$$

$$= \frac{2}{\pi}\left[\tan^{-1}x\right]_0^{\infty}$$

$$= \frac{2}{\pi} \times \frac{\pi}{2} = 1$$

E9) Suppose X denotes the number of minutes past 8 that the passenger arrives at the bus stop. Since X is uniformly distributed, the density function of X is given by

$$f(x) = \begin{cases} \frac{1}{30} , & \text{if } 0 \leq x \leq 30 \\ 0 , & \text{otherwise} \end{cases}$$

(see Example 4)

Now the passenger will have to wait less than 5 minutes if and only if he or she arrives between 8.10 A.M. and 8.15 A.M. or between 8.25 A.M. and 8.30 A.M. Therefore the probability that the passenger waits less than 5 minutes is

$P[10 < X \leq 15] + P[25 < X \leq 30]$

But $P[10 < X \leq 15] = \int_{10}^{15} f(x)dx = \int_{10}^{15} \frac{1}{30}dx = \frac{1}{6}$

and $P[25 < X \leq 30] = \int_{25}^{30} \frac{1}{30}dx = \frac{1}{6}$.

Therefore the required probability $= \frac{1}{3}$.

E10) f cannot be a density function since $f(x) < 0$ for $\sqrt{2} < x < \frac{5}{2}$.

E11) i) Property (i) : Let f(x) denotes the density function of X.

From Definition 4, we have

$$E(Y) = E(aX + b) = \int_{-\infty}^{+\infty} (ax + b)\, f(x)dx$$

$$= a\int_{-\infty}^{+\infty} x\, f(x)dx + b\int_{-\infty}^{+\infty} f(x)dx$$

$$= a\,E(X) + b,$$

since $\int_{-\infty}^{+\infty} f(x)dx = 1$

Property V : Let $E(x) = \mu$. Then $E(aX + b) = a\mu + b$ by property (i). Therefore

$$Var(aX + b) = E[aX + b - a\mu + b]^2$$

$$= E[(aX - a\mu)^2]$$

$$= E\left[a^2 (X - \mu)^2\right]$$

$$= a^2 E[(x - \mu)^2]$$

$$= a^2\, Var(X).$$

Property VI :

$$Var(X) = E\,[(X - \mu)^2]$$

$$= E\left[X^2 - 2\mu x + \mu^2\right]$$

$$= E\,(X^2) + E\,[(-2\mu x)] + E\,\mu^2$$

$$= E(X^2) - 2\mu\, E(X) + \mu^2$$

$$= E(X^2) - 2\mu^2 + \mu^2$$

$$= E(X^2) - \mu^2.$$

E12) Expectation is zero and variance is 1/6.

E13) $E(X) = \int_1^{\infty} x\,\frac{2}{x^3}dx = \int_1^{\infty} \frac{2}{x^2}dx = \left[-\frac{2}{x}\right]_1^{\infty} = 2$

$$E(X^2) = \int_1^{\infty} x^2 \cdot \frac{2}{x} = \int_1^{\infty} \frac{2}{x}dx = 2[\ln x]_1^{\infty}$$

and $\ln x \to \infty$ as $x \to \infty$. Hence $E(X^2)$ is not finite and therefore the variance does not exist.

E14) Let $g(a) = E[(X-a)^2]$

$= E[(X-\mu+\mu-a)^2]$, where $\mu = E(X)$

$= E[(X-\mu)^2] + (\mu-a)\,E[(X-\mu)] + (\mu-a)^2$

$= E[(X-\mu)^2] + (\mu-a)^2$, since $E[(X-\mu)] = 0$.

This shows that g(a) will be minimum when $a = \mu = E(X)$.

E15) $$E(x^r) = \int_\alpha^\beta x^r \frac{1}{\beta-\alpha}dx$$

$$= \frac{1}{\beta-\alpha}\left[\frac{x^{r+1}}{r+1}\right]_\alpha^\beta = \frac{\beta^{r+1}-\alpha^{r+1}}{(r+1)(\beta-\alpha)}$$

In particular

$$\mu = E(X) = \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} = \frac{\beta+\alpha}{2}$$

Further

$$E(X-\mu)^r = E\left[\sum_{j=0}^{r}\binom{r}{j}x^j(-\mu)^{r-j}\right]$$

$$= \sum_{j=0}^{r}\binom{r}{j}(-\mu)^{r-j}E(X^j)$$

$$= \sum_{j=0}^{r}\binom{r}{j}(-\mu)^{r-j}\frac{(\beta^{j+1}-\alpha^{j+1})}{(r+1)(\beta-\alpha)}$$

where $\mu = \frac{\alpha+\beta}{2}$.

E16) By definition, the skewness is given by

$$r_1 = \frac{E[(X-\mu)^3]}{\sigma^3}$$

Then show that $E(X) = 1$, $E(X^2) = 2$ and $E(X^3) = 6$.

Hence $\sigma^2 = 1$ and $\mu = 1$.

Hence we have

$\gamma_1 = E[(X-1)^3]$

$= E[X^3 - 3X^2 + 3X - 1]$

$= 2$.

E17) Let $E(X) = \mu_X$, $Var(X) = \sigma_X^2$. Denote the coefficients of skewness and kurtosis of X by ${\gamma_1}^X$ and ${\gamma_2}^X$ respectively. Then $\mu_Y = E(Y) = a\,\mu_X + b$ and $\sigma_Y^2 = a^2\sigma_X^2$.

Further

$${\gamma_1}^Y = \frac{E[(Y-\mu Y)^3]}{\sigma_Y^3} = \frac{E[(aX+b)-(a\mu_X+b)]^3}{a^3\sigma_X^3}$$

$$= a^3\frac{E[(X-\mu_X)^3]}{a^3\sigma_X^3} = \frac{E[(X-\mu_X)^3]}{{\sigma_X}^3} = {\gamma_1}^X$$

Similarly we can show that ${\gamma_2}^Y = {\gamma_2}^Y$. Hence the result

E18) The first moment $m_1 = E(X) = \frac{d}{dt}M_X(t)\big|_{t=0}$

Here $$M_X(t) = \frac{e^{t\beta}-e^{t\alpha}}{t(\beta-\alpha)}$$

$$E(X) = \frac{d}{dt}\left(\frac{e^{t\beta}-e^{t\alpha}}{t(\beta-\alpha)}\right)\Big|_{t=0}$$

$$= \frac{1}{\beta-\alpha}\left[\frac{t(\beta e^{t\beta}-\alpha e^{t\beta})-(e^{t\beta}-e^{t\alpha})}{t^2}\right]_{t=0}$$

But when we substitute t = 0, the expression the R.H.S. is of the $\frac{0}{0}$ form. Therefore, by applying, L'hopital's rule for $\frac{0}{0}$ form (see MTE–07, Block 1, Unit 2), we get

$$E(X) = \frac{1}{\beta - \alpha}\left[\frac{(\beta e^{t\beta} - \alpha e^{t\alpha} + \beta^2 e^{t\beta} - \alpha^2 e^{t\alpha}) - (\beta e^{t\beta} - \alpha e^{t\alpha})}{2t}\right]_{t=0}$$

$$= \left[\frac{\beta^2 e^{t\beta} - \alpha^2 e^{t\alpha}}{2(\beta - \alpha)}\right]_{t=0}$$

$$= \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{\beta + \alpha}{2}.$$

Similar argument shows that

$$Var(X) = \frac{(\beta - \alpha)^2}{12}.$$

E19) $M_X(t) = E(e^{tX}) = \int_{-\infty}^{+\infty} e^{tx} f(x)dx$

$$= \int_0^1 e^{tx} x\,dx + \int_0^2 e^{tx}(2 - x)dx$$

$$= \left[\frac{e^t - 1}{t}\right]_0^2, \quad t \neq 0$$

and $M_X(0) = 1$.

E20) $M_X(t) = \frac{1}{(1 - t)^2}, t < 1$

E21) $M_Y(t) = E\left[e^{t(aX + b)}\right]$

$$= E\left[e^{taX} e^{tb}\right]$$

$$= e^{tb} E\left[e^{taX}\right]$$

$$= e^{tb} M_X(at)$$

E22) Let $Y = g(X) = X^2$ and f_Y denote the density function of Y. Then

$$f_Y(y) = f_X(s(y)) \frac{d}{dy} s(y)$$

where f_X denotes the density function of X and s(y) denotes the inverse function of g(x).

$$\therefore \quad f_Y(y) = f_X\left(y^{1/2}\right) \frac{d}{dy}\left(y^{1/2}\right), \quad 0 < y < 1$$

$$= 1 \times \frac{1}{2} y^{-1/2}$$

$$= \frac{1}{2\sqrt{y}}, \quad 0 < y < 1$$

$$\therefore \quad f_Y(y) = \frac{1}{2\sqrt{y}}, \quad \text{if } 0 < y < 1$$

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

E23) $f_Y(y) = \frac{1}{6(4 - y)^{1/3}}$.

UNIT 11 STANDARD CONTINUOUS DISTRIBUTIONS

Structure

11.1 INTRODUCTION

In the previous unit, we have introduced the notions of distribution function, density function, expectation, variance and moments for a univariate continuous random variable. You have seen several examples dealing with these concepts. In this unit, we study the properties of some standard (absolutely) continuous distributions. These distributions are widely used in statistical inference as you will see in Block 4. Our main emphasis here is on normal, exponential, gamma and beta distribution. We have chosen these distributions because of their wide spread applicability. You have already met some of these in Unit 10. Here we shall take them one by one and discuss them in detail.

Objectives

- After reading this unit, you should be able to :
- compute the mean and variance of the normal, exponential, gamma and beta distributions ;
- investigate properties of other distributions which you come across.

Let us start with the normal distribution.

11.2 NORMAL DISTRIBUTION

Normal distribution, also called Gaussion distribution, is the most important probability distribution. It was found that the normal distribution is a good fit for a large class of data sets found in practice. For instance, normal distribution is a good approximation to the distribution of heights of people in a particular region or to the distribution of marks obtained by students from a university in a particular examination or to the distribution of diameters of bolts produced in a certain factory. It has been emphatically observed that normal distribution is also a good fit for the distribution of measurement and was derived by Gauss under certain assumptions as the probability law governing such errors.

Another reason for the importance of normal distribution in Statistics is the **central limit theorem.** We will discuss more about this theorem in Unit 14. What it essentially implies is that, even though the original set of observations might not be

from a normal distribution, the averages of these observations will be distributed approximately normal as long as the number of observations is large. You will realise the significance of this statement a little later.

Let us first look at the definition of normal distribution.

In Example 9 of Unit 10 we had considered the special case when $\mu = 0$ and $\sigma = 1$.

A random variable X is said to have **normal distribution** if it has a probability density function (p.d.f.) f of the form

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

where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$ are certain fixed quantities referred to as population parameters. Let us now check whether the function f is in fact a density function. Recall that a function f is a density function if and only if $f(x, \mu, \sigma^2) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f(x; \mu, \sigma^2)dx = 1.$$

Non-negativity of f is obvious from its definition. Let us check the second condition :

$$\int_{-\infty}^{\infty} f(x; \mu, \sigma^2)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\, e^{-y^2/2}} dy, \text{ (by the transformation } y = \frac{x-\mu}{\sigma})$$

Now, to evaluate this integral, we start by writing

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

Then

$$I^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\, e^{-x^2/2}} dz$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}[y^2 + z^2]\right\} dy\, dz.$$

Apply the transformation $y = r\cos\theta$ and $z = r\sin\theta$. Then, from the change of variable formula for double integrals (see Unit 11, Block 4, MTE -07), if follows that

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} \exp\left\{-\frac{1}{2}r^2\right\} r\, dr\, d\theta$$

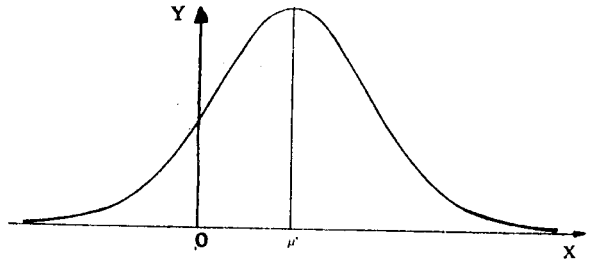
$$= \frac{1}{2\pi} 2\pi \int_0^{\infty} r \exp\left\{-\frac{1}{2}r^2\right\} dr = 1$$

Hence I = 1.

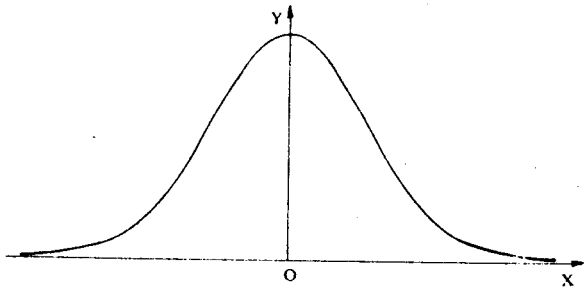
This proves that $\int_{-\infty}^{\infty} f(x; \mu, \sigma^2)\, dx = 1.$

You know that if $\mu = 0$ and $\sigma = 1$, then the distribution is called **standard normal distribution**. We usually denote the standard normal distribution by $\Phi(x)$ and the

standard normal density by $\varphi(x)$. In Fig. 1 you can see graphs of the p.d.f. of the standard normal distribution and that of an arbitrary normal distribution.



(a)



(b)

Fig. 1 : The graph of the p.d.f. of a (a) normal distribution (b) standard normal distribution.

From Fig. 1(a) you can see that the density function $f(x; \mu, \sigma^2)$ is symmetric about μ. Further μ is the median as well as the mode of the distribution. The graph of f is bell shaped. The shaded area in Fig. 1(b) gives the standard normal distribution function. Now let us calculate the expectation, variance and moments for a normal distribution.

Example 1 : Let us calculate the expected value and the variance of the r.v. X, whose density function, f, is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty.$$

where $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$.

Haven't you seen a similar density function before? In Unit 10, Example 9 you have seen a particular case of this with $\mu = 0$ and $\sigma = 1$.

Now,

$$E(X) = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} (\sigma y + \mu) \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy, \text{ if we put } \frac{x-\mu}{\sigma} = y$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2} dy + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

$$= \frac{\sigma}{\sqrt{2\pi}} \left[e^{-y^2/2}\right]_{-\infty}^{\infty} + \mu, \text{ since} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = 1$$

$$= \mu.$$

Also, Var(X) = E[(X − μ)²]

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$$

$$= \sigma^2 \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy, \text{ by putting } y = \frac{x-\mu}{\sigma}$$

Solve this integral by using the method of integration by parts, and check that

$$\int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = 1.$$

Using this we get

$$\text{Var}(X) = \sigma^2.$$

Next we shall find the moments.

Example 2 : Suppose X is a random variable with density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty.$$

Then

$$\mu_k = E[(x-\mu)^k]$$

$$= \int_{-\infty}^{\infty} (x-\mu)^k \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx$$

$$= \sigma^k \int_{-\infty}^{\infty} y^k \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

If k is an odd integer, then the integrand is an odd function and hence

$$\int_{-\infty}^{\infty} y^k e^{-y^2/2} dy = 0$$

This implies that

$$\mu_k = 0, \text{ if k is odd.}$$

Now suppose k is even. Let k = 2m where m is a positive integer.

Then

$$\mu_{2m} = \frac{2\sigma^{2m}}{\sqrt{2\pi}} \int_0^{\infty} y^{2m} e^{-y^2/2} dy$$

$$= \frac{2\sigma^2}{\sqrt{2\pi}} \int_0^{\infty} (2z)^m e^{-z} (2z)^{-1/2} dz, \text{ by putting } z = \frac{y^2}{2}$$

$$= \frac{1}{\sqrt{2}} \frac{2^{m+1}\sigma^{2m}}{\sqrt{2\pi}} \int_0^{\infty} z^{m-1/2} e^{-z} dx$$

$$= 2^m \frac{\sigma^{2m}}{\sqrt{\pi}} \Gamma\left[m + \frac{1}{2}\right].$$

Recall that we have seen the gamma function

$$\Gamma(a) = \int_0^{\infty} z^{\alpha-1} e^{-z} dz.$$

In Unit 10.

Now we make use of some more properties of the gamma function.

$\Gamma(1) = 1$

$\Gamma(1/2) = \sqrt{\pi}$,

$\Gamma(\alpha + 1) = \alpha\,\Gamma(\alpha)$ for any $\alpha > 0$,

$\Gamma(k + 1) = k!$, if k is a non-negative integer.

and

$$\frac{\Gamma(m)\,\Gamma(n)}{\Gamma(m+n)} = \int_0^1 u^{m-1}(1-u)^{n-1}\,du.$$

Using these properties you can check that

$$\Gamma\left[m + \frac{1}{2}\right] = \left[\frac{1}{2} + (m-1)\right]\left[\frac{1}{2} + (m-2)\right] \cdots \left[\frac{1}{2}\right]\sqrt{\pi}$$
$$= \frac{(2m-1)(2m-3)\ldots\ldots 1}{2^m}\sqrt{\pi}.$$

Hence,

$$\mu_{2m} = [(2m-1)(2m-3)\ldots\ldots 1]\,\sigma^{2m}, \quad m \geq 1.$$

In particular, we get $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$, and hence,

$$\mu_4 = 3\,\mu_2^2.$$

For the normal density function, we also have

$$\gamma_1 = \frac{E[(X-\mu)^3]}{\sigma^3} = 0 \quad \text{and} \quad \beta_2 = \frac{E[(X-\mu)^4]}{\sigma^4} = 3$$

$\gamma_1 = 0$ implies that the normal density is symmetric. And $\beta_2 = 3$ implies that it is meso- kurtic (see Unit 3, Block 1).

An in the previous unit you must have realized that the calculation of moments is a somewhat tedious process. Can you guess what an easier way is? This is what we shall do now.

Example 3 : Suppose X has the normal density function with mean μ and variance σ^2. Then

$$M_X(t) = E(e^{tX})$$

$$= \int_{-\infty}^{\infty} e^{tx}\,\frac{1}{\sqrt{2\pi}\,\sigma^2}\,e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}\,dx$$

$$= \int_{-\infty}^{\infty} e^{t(\sigma y + \mu)}\,\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}y^2}\,dy \qquad \text{(by the transformation } \frac{x-\mu}{\sigma}$$

$$= e^{\mu t}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\,e^{\frac{-1}{2}y^2 + t\sigma y}\,dy$$

$$= e^{\mu t}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\,e^{\frac{-1}{2}(y-t\sigma)^2 + \frac{t^2\sigma^2}{2}}\,dy$$

$$= e^{\mu t + \frac{1}{2}t^2\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(y-t\sigma)^2}\,dy.$$

The integral ... since the integrand can be considered as the density function of a normal distribution with mean $t\alpha$ and variance 1.

Hence,

$$M_X(t) = e^{\mu t + \frac{1}{2}t^2\sigma^2}, \quad -\infty < t < \infty.$$

Thus in the case of normal distribution, the m.g.f. exists for all real t.

We now state and prove an important theorem. You will realise the usefulness of this theorem from the examples that follow.

Theorem 1: If a random variable x has normal distribution with mean μ and variance σ^2, then the random variable $Y = aX + b$ has normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$.

Proof : Let us compute the m.g.f. or Y. Then

$$\begin{aligned} M_Y(t) &= E\left[e^{tY}\right] \\ &= E\left[e^{t(aX+b)}\right] \\ &= e^{tb}E\left[e^{taX}\right] \\ &= e^{tb}M_X(at) \\ &= e^{tb}.e^{\mu at + \frac{1}{2}a^2\sigma^2t^2} \\ &= e^{(a\mu+b)t + \frac{1}{2}a^2\sigma^2t^2} \end{aligned}$$

for $-\infty < t < \infty$.

Let Z be a random variable with normal distribution having mean $a\mu + b$ and variance $\frac{1}{2}a^2\sigma^2$.

Then $$M_Z(t) = e^{(a\mu+b)t + \frac{1}{2}a^2\sigma^2t^2}, \quad -\infty < t < \infty.$$

Since $M_Y(t) = M_Z(t)$ for all t, and in particular in a neighbourhood of zero, it follows from Theorem 1 in Unit 10, that the distribution of Y and Z are the same. Hence the distribution of Y is normal with mean $a\mu + b$ and variance $a^2\sigma^2$.

For simplicity, we denote the normal distribution with mean μ and variance σ^2 by $N(\mu, \sigma^2)$. If follows, from Theorem 1 that if Y is $N(\mu, \sigma^2)$, then $X = \frac{Y-\mu}{\sigma}$ is $N(0, 1)$.

We now give an example to show how Theorem 1 can be used to find the probabilities relating to an r.v. with normal distribution.

Example 4 : Suppose Y is N(1, 4). Let us compute $P[3 < Y < 5]$.

Here $\mu = 1$ and $\sigma = 2$. We apply the transformation $X = \frac{Y-1}{2}$.

Then we obtain that X is N(0, 1) and

$$P[3 < Y < 5] = P\left[\frac{3-1}{2} < X < \frac{5-1}{2}\right] = P[1 < X < 2]$$

Now

$$P[1 < X < 2] = \Phi(2) - \Phi(1) = \int_1^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx$$

where Φ denotes the standard normal distribution function.

Thus, we have found the required probability relating to the variable Y in terms of the probabilities relating to X, which has a standard normal distribution. So, if we know the probabilities relating to a standard normal distribution, we will be able to calculate those for any normal distribution. But in general, the exact evaluation of $\Phi(x)$ is not possible since there is no explicit formula for computing definite integral.

$$\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx$$

on the desk calculator.

However, extensive tables giving probabilities calculated from the standard normal distribution are available (see Table 1 in Appendix). We can use these tables for computing probabilities connected with a general normal distribution. You may find the computation easy if you not that

$$\Phi(-x) = 1 - \Phi(x) \; \forall \; x.$$

This is because by the symmetry of the standard normal density function about zero, the area to the left of x must be equal to (1 - area of the right of x).

Let us now see how we use Table 1 in the Appendix. From Table 1, check that

$$\Phi(3.00) = 0.9987, \text{ and}$$

$$\Phi(2.00) = 0.9773$$

Hence, $P[-3 < X < 3] = \Phi(3) - \Phi(-3)$

$$= \Phi(3) - [1 - \Phi(3)]$$

$$= 2\Phi(3) - 1$$

$$= 0.9974$$

Similarly,

$$P[-2 < X < 2] = 0.95546.$$

Now, how do we interpret this? We can say that more than 99% of the total probability is carried by the interval $[-3, 3]$ and about 95% of the total probability is supported by the interval $[-2, 2]$ for a standard normal distribution.

In general, if Y is $N(\mu, \sigma^2)$, then using the transformation $X = \frac{Y - \mu}{\sigma}$, we get

$$P[\mu - 3\sigma < Y < \mu + 3\sigma] = P[-3 < X < 3] = 0.9974$$

and $P[\mu - 2\sigma < Y < \mu + 2\sigma] = .9546.$

Now if you go back to Example 1, then using Table 1 we get

$$P[1 < X < 2] = \Phi(2) - \Phi(1) = 0.9772 - 0.8413 = 0.1359.$$

Then $P[3 < Y < 5] = P[1 < X < 2] = 0.1359.$

Now, we'll show how the normal distribution is used in some practical situations.

Example 5 : It has been observed that the marks obtained by the students in an examination generally follows a normal distribution. In other words the approximating curve of the histogram obtained from the data should be bell-shaped (see Fig. 2).

This method is referred to as 'grading on the curve'.

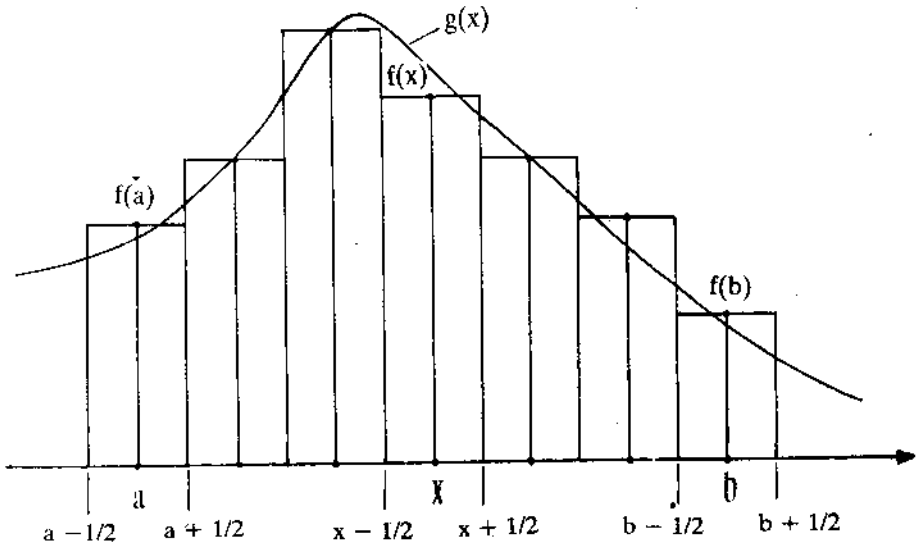


Fig. 2 : The histogram is approximated by a normal curve.

The teacher uses the percentage marks to estimate the mean μ and the variance σ^2. Suppose the letter grade A is assigned to a student if the percentage of marks is greater than $\mu + \sigma$, B to those whose percentage is between μ and $\mu + \sigma$, C to those between $\mu - \sigma$ and μ, D to those between $\mu - 2\sigma$ and $\mu - \sigma$ and E to those getting a percentage below $\mu - 2\sigma$. Suppose we want to calculate the percentage of students having grades A, B, C, D and E.

Let X denote the r.v. corresponding to the data and μ and σ denote the mean and standard deviation respectively. Then the percentages of students having grades A, B, C, D and E are given by $P[x > \mu + \sigma]$, $P[\mu < X < \mu + \sigma]$, $P[\mu - \sigma < X < \mu]$, $P[\mu - 2\sigma < X < \mu - \sigma]$ and $P[X < \mu - 2\sigma]$ respectively. Applying the transformation $Y = \frac{X - \mu}{\sigma}$, we find that Y has normal distribution, N(0, 1). Then

$$P[X > \mu + \sigma] = P\left[\frac{X - \mu}{\sigma} > 1\right]$$
$$= P[Y > 1]$$
$$= 1 - P[Y \leq 1]$$
$$= 1 - \Phi(1)$$
$$= 0.1587$$

$$P[\mu < X < \mu + \sigma] = P[0 < Y < 1]$$
$$= \Phi(1) - \Phi(0)$$
$$= 0.3413.$$

Similarly we get

$$P[\mu - \sigma < X < \mu] = \Phi(0) - \Phi(-1) = 0.3413.$$

$$P[\mu - 2\sigma < X < \mu - \sigma] = \Phi(-2) - \Phi(-1) = 0.1359.$$

and

$$P[X < \mu - 2\sigma] = \Phi(-2) = 0.0228.$$

In other words, approximately 16% will receive grade A, 34% grade B, 34% grade C, 14% grade D and 2% will fail, since they get grade E.

Now, try to solve these exercises.

E1) If X is N (0, 1), find

a) $P[0 \le X \le 0.87]$

b) $P[-2.64 \le X \le 0]$

c) $P[-2.13 \le X \le -0.56]$

d) $P[|X| > 1.39]$.

E2) If the m.g.f. of X is $M_X(t) = \exp(-6t + 32t^2)$, find $P[-4 \le X < 16]$.

E3) Show that if X has a normal distribution with zero mean, so does –X.

E4) The height of female students at IGNOU follows approximately a normal distribution, with mean 60 inches and standard deviation 2. Find the probability that a female student selected at random has height

a) less than 58 inches

b) between 58 inches and 62 inches.

E5) If X is $N(\mu, \sigma^2)$, determine the density function of $Y = \frac{(X-\mu)^2}{\sigma^2}$.

E6) A random variable X is said to have log-normal distribution if ln X has normal distribution. Suppose ln X is $N(\mu, \sigma^2)$.

Find the density of X.

Show that

$E(X) = \exp[\mu + \frac{\sigma^2}{2}]$, and

$\text{Var}(X) = \{\exp(\sigma^2) - 1\} \{\exp(2\mu + \sigma^2)\}$

We now consider two more distributions which are found quite useful in applications. You have already met one of these before in Unit 10 (see Example 8).

11.3 EXPONENTIAL AND GAMMA DISTRIBUTIONS

Let us quickly recall the facts you have already studied in Unit 10 about the exponential distribution.

11.3.1 Exponential Distribution

From Unit 10 you know that a random variable X is said to be exponentially distributed if it has a density function of the form

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0. \end{cases}$$

where $\lambda > 0$. You know that the distribution function F, corresponding to the density f is

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } x \ge 0 \\ 0, & \text{if } x < 0. \end{cases}$$

Further, $E(X) = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$ (See Example 8 of Unit 10).

The exponential distribution serves as a good model whenever there is a waiting time involved for a specific event to occur. For example the moment of waiting time

(say, starting ... ... an accident to occur, or for a telephone call to be received, follows an exponential distribution under reasonable conditions.

An important property of an exponential distribution is its **memoryless** property. In other words,

$$P[X > s + t \mid X > t] = P[X > s] \quad \ldots(1)$$

for all $s, t \geq 0$. This follows from the following observation

P (A | B) denotes the conditional probability of an event A given the event B.

$$P[X > s + t \mid X > t] = \frac{P[X > s + t]}{P[X > t]} = \frac{1 - P[X \leq s + t]}{1 - P[X \leq t]}$$

$$= \frac{1 - (1 - e^{-(s+t)x})}{1 - (1 - e^{-tx})}$$

$$= \frac{e^{-(s+t)x}}{e^{-tx}}$$

$$= e^{-sx}$$

$$= P[X > s].$$

Now we shall illustrate why this property is called "Forgetfulness".

If we interpret X as the life time of a component (say a light bulb) in hours, the above equation, (1) states that the probability that the bulb lasts for at least s + t hours given that it has worked for at least t hours is the same as the probability that it lasts for s hours. That is, if the bulb is working after t hours, then the distribution of the remaining time that it works is the same as the original distribution. In other words, the bulb **does not remember** that it has already been in use for t hours.

Recall that we have noted in Unit 9 that a discrete r.v. with geometric distribution also has a similar forgetfulness property.

We can show that the exponential distribution is the one and only one continuous distribution which has the forgetfulness property.

This means that the only continuous r.v. X assuming non-negative values for which $P[X > s + t \mid x > s] = P[X > t]$, for all $s, t \geq 0$, is an exponentially distributed r.v. Here is how we go about it.

Suppose X is a non-negative random variable such that

$$P[X > s + t \mid X > t] = P[x > s], \text{ i.e. } \frac{P[X > s + t]}{P[X > t]} = P[X > s]$$

Let $\bar{F}(x) = P[X > x]$. Then the above equation reduces to

$$\bar{F}(s + t) = \bar{F}(s)\,\bar{F}(t)$$

for all $s, t \geq 0$. Now you can check that the only right continuous solution of this equation is

$$\bar{F}(s) = e^{-\lambda s}, s \geq 0$$

or equivalently,

$$F(s) = 1 - e^{-\lambda s}, s \geq 0.$$

(see E7).

Let us consider an example.

Example 6 : Suppose that accidents occur in a large industrial plant at a rate of $\lambda = \frac{1}{10}$ per day. Suppose we begin observing the occurrence of these accidents at the starting of work on Monday. Let X be the number of days until the first accident occurs. Then the distribution of X is

$$F(x) = 1 - e^{-\frac{1}{10}x}, x \geq 0$$
$$= 0, x < 0.$$

The probability that the first week is accident free is

$$P[X > 5] = e^{-5/10} = e^{-1/2}$$
$$= 0.6065$$

[This value is obtained by using a scientific calculator.]

The probability that the first accident occurs on wednesday of the second week is

$$P[7 < X \leq 8] = [1 - e^{-8/10}] - [1 - e^{-7/10}]$$
$$= 0.047.$$

(Here also we have used a scientific calculator to compute the values.)

You can now try these exercises.

E7) Let F be the distribution function of a non-negative random variable X. Show that the only solution of the functional equation

$$\bar{F}(s + t) = \bar{F}(s)\,\bar{F}(t), t \geq 0, s \geq 0$$

is

$$F(s) = 1 - e^{-\lambda s}, s > 0$$
$$= 0 \quad , s \leq 0$$

for some $\lambda > 0$.

E8) Calls arrive at a switchboard following an exponential distribution with parameter $\lambda = 5$ per hour. If we are at the switchboard, what is the probability that the waiting time for a call is

i) at least 15 minutes,

ii) not more than 10 minutes,

iii) exactly 5 minutes.

We have seen in Unit 8 that the Poisson distribution can be chosen as a model for the distribution of the number of occurrences of an event during a fixed time interval. Let us denote by N(t) the number of occurrences of the event (such as a telephone call, or an accident etc.) in the interval [0, t] and assume that, for $t > 0$, N(t) has Poisson distribution with mean λt, where $\lambda > 0$. λ denotes the average rate of occurrence in unit interval. Assume that $N(0) = 0$.

Let us denote by X the waiting time involved before the first occurrence. It is clear that

$$P[X > t] = P[N(t) = 0]$$
$$= e^{-\lambda t}, \text{ for } t > 0.$$

Therefore,

$$P[X \leq t] = 1 - e^{-\lambda t}, \text{ if } t > 0$$
$$= 0, \text{ if } t < 0.$$

Thus X has an exponential distribution. These relations show the link between the exponential and the Poisson distributions. So you know that the exponential distribution describes the probability distribution of the waiting time for the **first occurrence**. Now, suppose we are interested in the waiting time for **r occurrences**. Let us denote by Y the waiting time that elapses before r occurrences of the event.

Then,

$$P[Y > t] = P[N(t) < r]$$

$$= \sum_{j=0}^{r-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}$$

or equivalently the distribution function of Y is given by

$$F_Y(t) = P[Y \leq t] = 1 - \sum_{j=0}^{r-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}, \; t > 0$$

Obviously $F_Y(t) = 0$ for $t \leq 0$.

We can now obtain the density function of Y by differentiating $F_Y(t)$ with respect to t. You should check that

$$f_Y(t) = \begin{cases} \frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t} & , \quad t > 0 \\ 0 & , \quad t \leq 0. \end{cases}$$

In the above discussion, r is an integer greater than or equal to 1. Note that if r = 1, this density function reduces to the exponential density studied earlier.

The density function given above is a special case of the gamma density function which we shall discuss in the next section.

11.3.2 Gamma Distribution

In the last unit and in sub-sec. 11.2.1 of this unit, you were introduced to a new function known as the 'gamma function'. Before defining the gamma distribution, let us first recall the definition of the gamma function.

For any positive number α, the gamma function, denoted by $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx$$

(The integral on the right hand side exists for any $\alpha > 0$.) You have to already come across some properties of the gamma function in Sec. 11.2 and used these properties for evaluating certain integrals.

Now with the help of the gamma function we shall define a gamma distribution.

Let X be a random variable with density function,

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & , \quad x > 0 \\ 0 & , \quad x \leq 0 \end{cases}$$

where $\alpha > 0$ and $\lambda > 0$ and Γ is the gamma function. Then X is said to have **gamma distribution with parameters α and λ.**

Check that f(x) satisfies the properties of a density function. Fig. 3(a) and (b) show graphs of some typical gamma density functions.

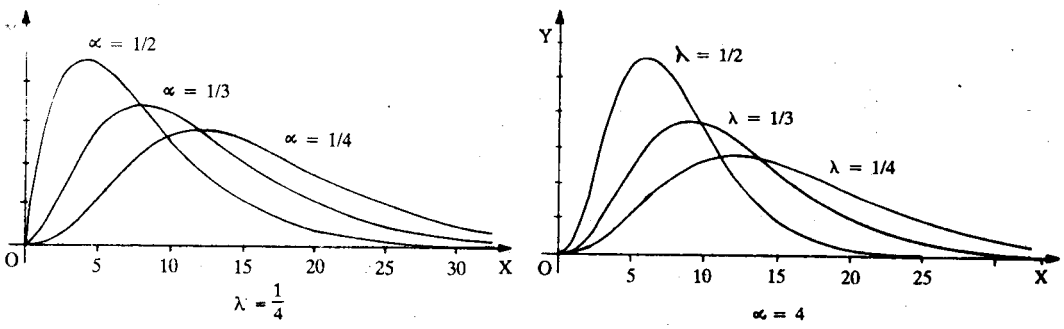


Fig. 3 : Graphs of Gamma Density Functions for different values of (a) α (b) λ.

Let us compute the m.g.f. of the gamma density function now. By definition,

$$M_X(t) = E[e^{tx}] = \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\lambda - t)x} dx$$

The last integral is finite if and only if $\alpha > 0$ and $\lambda - t > 0$. This can be proved by using some properties of Γ. But for the time being you will have to take our word for it. Hence the m.g.f. $M_X(t)$ exists if and only if $t < \lambda$ $\alpha > 0$. Then, for $t < \lambda$,

$$M_X(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left[\frac{y}{\lambda - t}\right]^{\alpha-1} \frac{e^{-y}}{(\lambda - t)} dy$$

$$= \frac{\lambda^\alpha}{(\lambda - t)^\alpha \Gamma(\alpha)} \int_0^\infty y^{\alpha-1} e^{-y} dy$$

$$= \frac{\lambda^\alpha}{(\lambda - t)^\alpha}.$$

Check that

$$\frac{dM_X(t)}{dt} = \lambda^\alpha(-\alpha)(\lambda - t)^{-\alpha-1}(-1)$$

$$= \alpha \lambda^\alpha (\lambda - t)^{-\alpha-1}$$

and

$$\frac{d^2M_X(t)}{dt^2} = \alpha(\alpha+1)\lambda^\alpha (\lambda - t)^{-\alpha-2}$$

In particular

$$E(X) = \frac{dM_X(t)}{dt}\bigg|_{t=0} = \frac{\alpha \lambda^\alpha}{\lambda^{\alpha+1}} = \frac{\alpha}{\lambda}$$

and

$$E(X^2) = \frac{\alpha(\alpha+1)\lambda^\alpha}{\lambda^{\alpha+2}} = \frac{\alpha(\alpha+1)}{\lambda^2}$$

Hence

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

$$= \frac{\alpha}{\lambda^2}.$$

The gamma distribution is a good model for waiting times. It has also been used as a model for the distribution of incomes as the parameters α and λ provide a flexibility in fitting the model to the data. The following exercises will lead to a better understanding of this distribution.

E9) Let X be a gamma random variable with parameters α and λ. Compute $E(X^m)$ for all $m \geq 1$ directly, and hence derive E(X) and Var(X).

E10) Suppose X is a gamma random variable with E(X) = 2 and Var(X) = 7. Find α and λ.

E11) Suppose X is N(0, 1). Show that $Y = X^2$ has gamma distribution with $\alpha = 1/2$ and $\lambda = 1/2$.

In the next section we shall take up one last distribution the beta distribution.

11.4 BETA DISTRIBUTION

While studying Statistics and many other sciences, scientists found that they came across some particular functions quite often. They have identified a few such functions. These functions are called **special functions.** You have already seen one special function so far – the gamma function. You have also seen how the gamma distribution is defined with the help of this function. Now we are going to introduce another special function, the beta function, B. Then, with the help of B we shall define the beta distribution.

If $\alpha > 0$, $\beta > 0$, then

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \, dx$$

is called the **beta function.**

The beta function is related to the gamma function in the following manner.

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} \text{ for all } \alpha > 0, \beta > 0.$$

Let's prove this.

From the definition of the gamma function, we get

$$\Gamma(\alpha)\,\Gamma(\beta) = \left\{\int_0^\infty x^{\alpha-1} e^{-x} \, dx\right\} \left\{\int_0^\infty y^{\beta-1} e^{-y} \, dy\right\}$$

$$= \int_0^\infty \int_0^\infty x^{\alpha-1} y^{\beta-1} e^{-(x+y)} \, dx \, dy$$

Now to evaluate the integral we apply the transformation

$$u = \frac{x}{x+y} \text{ and } v = x+y.$$

You will find discussion of Jacobians and double integrals in Blocks 3 and 4 of MTE-07.

Then what are x and y in terms of u and v ?

$$x = u\,v \text{ and } y = (1-u)v$$

and the Jacobian of the transformation is

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

$$= \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v(1-u) + uv = v.$$

Note that as x and y take values in $[0, \infty]$, u varies over $[10, 1]$ and v varies over $[0, \infty]$

Hence, from the change of variable formula for double integrals, it follows that

$$\Gamma(\alpha)\Gamma(\beta) = \int_0^1 \int_0^\infty u^{\alpha-1} (1-u)^{\beta-1} v^{\alpha+\beta-1} e^{-v} \, dv du$$

$$= \int_0^\infty v^{\alpha+\beta-1} e^{-v} \, dv \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} \, du$$

$$= \Gamma(\alpha+\beta) \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} \, du.$$

$$= \Gamma(\alpha+\beta) \, B(\alpha, \beta).$$

This proves the required relation.

Now, we say that a r.v. X has **beta distribution** with parameters α and β, if X has the density function f, given by

$$f(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & , \quad 0 < x < 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

where $\alpha > 0, \beta > 0$.

An alternative representation for a beta density function with parameters α and β is

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, & 0 < x < 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Let us now compute the moments of a random variable X with the beta distribution having parameters α and β. For any integer $k \geq 1$.

$$E[X^k] = \int_0^1 x^k f(x) \, dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^1 x^k \cdot x^{\alpha-1} (1-x)^{\beta-1} \, dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^1 x^{k+\alpha-1} (1-x)^{\beta-1} \, dx$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} B(k+\alpha, \beta)$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \frac{\Gamma(k+\alpha)\,\Gamma(\beta)}{\Gamma(k+\alpha+\beta)}$$

$$= \frac{\alpha(\alpha+1)....(\alpha+k-1)}{(\alpha+\beta), (\alpha+\beta+1)....(\alpha+\beta+k-1)},$$

In particular by choosing k = 1 and k = 2, we have

$$E(X) = \frac{\alpha}{\alpha+\beta} \text{ and } E(X^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

Therefore,

$$Var(X) = E(X^2) - (E(X))^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Remark 1 : If $\alpha = 1$ and $\beta = 1$, then the beta distribution reduces to the uniform distribution on [0, 1].

In Fig. 4 you can see some typical graphs of beta density function.

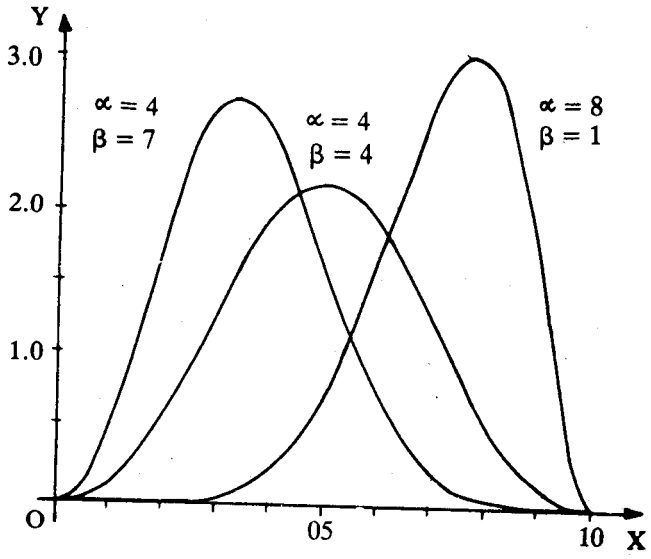


Fig. 4 : Graphs of B(α, β) for different values of α and β.

Now here are some exercises on the beta distribution.

E12) Evaluate the following numbers using the gamma function

(i) B(10, 7) (ii) $B\left(\frac{1}{2}, \frac{1}{2}\right)$ (iii) $B\left(\frac{5}{2}, \frac{7}{2}\right)$

E13) If X is a beta random variable with the parameters α and β, show that 1–X is a beta random variable with the parameters β and α.

E14) Determine the constant c such that the function

$$f(x) = cx^3(1-x)^6, \quad 0 < x < 1$$
$$= 0, \text{ otherwise}$$

is a density function.

This brings us to the end of this unit. Let us now summarise its main points.

11.5 SUMMARY

In this unit, we have derived the properties of the following distributions with emphasis on calculation of their distribution functions, density functions, their means, variances and moment generating functions :

1) Normal distribution :

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty$$

$$E(X) = \mu, \ Var(X) = \sigma^2.$$

If X is $N(\mu, \sigma^2)$, then $Y = \frac{X-\mu}{\sigma}$ is $N(0, 1)$

2) Exponential distribution :

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

$$E(X) = \frac{1}{\lambda}, \ Var(X) = \frac{1}{\lambda^2}.$$

This distribution is the only distribution with the forgetfulness property.

3) Gamma distribution :

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\mu(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$E(X) = \frac{\alpha}{\lambda}, \ Var(X) = \frac{\alpha}{\lambda^2}$$

The exponential and gamma distributions are also called **waiting time distribution**

4) Beta distribution:

$$f(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$E(X) = \frac{\alpha}{\alpha+\beta}, \ Var(X) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}.$$

11.6 SOLUTIONS AND ANSWERS

E1) i) From the Table given in the Appendix, we have

$$P[0 \leq X \leq 0.87] = F(0.87) - F(0)$$
$$= 0.8078 - 0.5000$$
$$= 0.3078$$

ii) $P[-2.64 \leq X \leq 0] = F(0) - F(-2.64)$

$$= 0.4959$$

iii) 0.2711

iv) $P[|X| > 1.39] = 1 - P[|X| \leq 1.39]$

$$= 1 - P[-1.39 \leq X \leq 1.39]$$

$$= 0.1646.$$

E2) From the m.g.f. of X, we get that X is a normal random variable with mean -6 and $\sigma = 8$ i.e. $N(-6, 64)$.

To calculate the probability $P[-4 \leq X \leq 16]$, we apply the transformation

$$Y = \frac{X+6}{8}$$

Then $P[-4 \le X < 16] = P\left[-\frac{4+6}{8} \le X < \frac{16+6}{8}\right]$

$= P[.25 \le X < 2.75]$

$= 0.3983$

E3) The m.g.f. of X is

$$M_X(t) = e^{0 \times t + \frac{1}{2}t^2\sigma^2} = e^{t^2\sigma^2/2}$$

Let $Y = -X$, since $E(X) = 0$, we have $E(-X) = 0$ and $Var(-X) = Var(X) = \sigma^2$. Then the m.g.f. of Y is

$$M_Y(t) = e^{0 \times t + \frac{1}{2}t^2\sigma^2}$$

$$= e^{t^2\sigma^2/2}$$

$$= M_X(t)$$

for $-\infty < t < \infty$. $\therefore$ by Theorem 1 in Unit 10, the distributions of X and $-X$ are the same. Hence $-X$ is $N(0, \sigma^2)$.

E4) Let $Y = \frac{X-60}{2}$

$$P[X < 58] = P\left[\frac{X-60}{2} < \frac{58-60}{2}\right]$$

$= P[Y < -1]$

$= \Phi(-1)$

$= 0.1587$

$P[58 < X < 62] = P[-1 < X < 1]$

$= \Phi(1) - \Phi(-1)$

$= 0.6826.$

E5) Suppose X is $N(\mu, \sigma^2)$. Let

$$Z = \frac{X-\mu}{\sigma} \text{ Then } Z^2 = Y = \frac{(X-\mu)^2}{\sigma^2}$$

Then, for $y > 0$ $P[Y \le y] = P[Z^2 \le y]$

$= P[-\sqrt{y} \le Z < \sqrt{y}]$

$$= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du.$$

Hence

$$f_Y(y) = 2\frac{1}{\sqrt{2\pi}} e^{-y/2} \cdot \frac{1}{2} y^{-1/2}$$

$$= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}, \quad y > 0$$

E6) $Y = \ln X$ is normal and $X = e^Y$. Hence

$E(X) = E(e^Y)$

and

$E(X^2) = E(e^{2Y})$

Since Y is $N(\mu, \sigma^2)$, $E[e^{tY}] = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

Therefore

$E(X) = E[e^Y] = e^{\mu + \frac{1}{2}\sigma}$

and

$E(X^2) = E[e^{2Y}] = 3^{2\mu} + 2\sigma^2$.

Hence

$$\begin{aligned} \text{Var}(X) &= e^{2\mu + 2\sigma^2} - \left(e^{\mu + \frac{1}{2}\sigma^2}\right)^2 \\ &= e^{2\mu + 2\sigma^2} - e^{2u + \sigma^2} \\ &= e^{2\mu + \sigma^2}\left[e^{\sigma^2} - 1\right]. \end{aligned}$$

E7) We prove a general result here. The only right continuous solution of the functional equation

$g(s+t) = g(s)g(t)$, $s > 0$, $t > 0$

which is not identically zero for $x > 0$, is

$g(s) = e^{-\lambda s}$, $\lambda > 0$.

This can be seen in the following way

Since $g(s+t) = g(s)\, g(t)$, it follows that

$$g\left(\frac{2}{n}\right) = g\left(\frac{1}{n} + \frac{1}{n}\right) = \left[g\left(\frac{1}{n}\right)\right]^2$$

and, in general, by repeated application, we have

$$g\left(\frac{m}{n}\right) = \left[g\left(\frac{1}{n}\right)\right]^m.$$

But

$$g(1) = g\left(\frac{1}{n} + \ldots\ldots + \frac{1}{n}\right) = \left[g\left(\frac{1}{n}\right)\right]^n.$$

Therefore

$$g\left(\frac{1}{n}\right) = [g(1)]^{1/n}$$

and

$$g\left(\frac{m}{n}\right) = [g(1)]^{m/n}.$$

By the right continuity of $g(x)$, it follows that

$$g(x) = [g(1)]^x.$$

for $-\infty < x < \infty$.

Note that $g(1) = \left[g\left(\frac{1}{2}\right)\right]^2 \geq 0$. If $g(1) = 0$, then $g(x) = 0$ for all $x > 0$ contradicting the fact that $g(x) \neq 0$ for $x > 0$. Hence $g(1) > 0$ and $g(x) = e^{-\lambda x}$, $x > 0$ where $\lambda = -\log g(1)$.

Since a distribution function $F(x)$ is right continuous, it follows that

$$\overline{F}(x) = e^{-\lambda x} \quad x > 0$$

or equivalently

$$F(x) = 1 - e^{-\lambda x},\ x > 0.$$

E8) The distribution is

$$F(X) = 1 - e^{-5x},\ x \geq 0$$

$$= 0,\ x < 0.$$

The probabilities are

i) $P[X \geq 15] = 0.2865$

ii) $P[X < 10] = 0.6565$

iii) $P[X = 5] = 0$.

E9) $$E(X^m) = \int_0^\alpha x^m \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\alpha x^{\alpha+m-1} e^{-\lambda x}\, dx$$

$$= \lambda^\alpha \int_0^\alpha x^{\alpha+m-1} e^{-\lambda x}\, dx$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(m+\alpha)}{\lambda^{m+\alpha}} = \frac{\Gamma(m+\alpha)}{\lambda^m \Gamma(\alpha)}$$

In particular

$$E(X) = \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha)}{\lambda\Gamma(\alpha)} = \frac{\alpha}{\lambda}$$

and

$$E(X^2) = \frac{\Gamma(\alpha+2)}{\lambda^2\Gamma(\alpha)} = \frac{(\alpha+1)\Gamma(\alpha+1)}{\lambda^2\Gamma(\alpha)}$$

$$= \frac{(\alpha+1)\alpha\,\Gamma(\alpha)}{\lambda^2\Gamma(\alpha)} = \frac{\alpha^2+\alpha}{\lambda^2}$$

Therefore

$$\text{Var}(X) = \frac{\alpha^2+\alpha}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}.$$

E10) Here $\frac{\alpha}{\lambda} = 2$ and $\frac{\alpha}{\lambda^2} = 7$. Hence $\lambda = \frac{2}{7}$ and $\alpha = \frac{4}{7}$.

E11) For $y > 0$, $P[X^2 \leq y] = P[-\sqrt{y} \leq X \leq \sqrt{y}]$

$$= 2 \int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du.$$

Hence the density function of Y is

$$f_Y(y) = 2 \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{1}{2} y^{-1/2}, \quad y > 0$$

$$= 0, \qquad y \leq 0.$$

This can be written in the form

$$f_Y(y) = \frac{\left(\frac{1}{2}\right)^{1/2}}{\Gamma(\frac{1}{2})} y^{(1/2)-1} e^{-y/2}, \quad y > 0$$

and

$$f_Y(y) = 0 \text{ for } y \leq 0$$

since $\Gamma(1/2) = \sqrt{\pi}$. This shows that the density function of Y is gamma density with $\alpha = 1/2$ and $\lambda = 1/2$.

E12) a) $B(10, 7) = \frac{9!6!}{16!} = 1.25 \times 0^{-5}$.

b) $B\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma(1)} = \Gamma\left(\frac{1}{2}\right)^2 = \pi$.

c) $B\left(\frac{5}{2}, \frac{7}{2}\right) = \frac{45\pi}{3840}$

E13) Let $Y = 1 - X$. Then $P[Y \leq y] = P[X \geq 1-y]$

Hence

$$F_Y(y) = 1 - F_X(1-y)$$

and

$$f_Y(y) = f_X(1-y).$$

Therefore

$$f_Y(y) = \frac{1}{B(\alpha, \beta)}(1-y)^{\alpha-1}\, y^{\beta-1}, \quad 0 < y < 1$$

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

which is the beta density with the parameters β and α.

E14) Here $\alpha = 4$ and $\beta = 7$. Further

$$C = \frac{1}{B(4, 7)} = \frac{\Gamma(4+7)}{\Gamma(4)\Gamma(7)} = \frac{\Gamma(1)}{\Gamma(4)\Gamma(7)} = \frac{10!}{3!6!} = 840.$$

Appendix

Table 1. Values of the Standard Normal Distribution Function

$$\phi(z) = \int_{-x}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = P(Z \leq z)$$

z	0	1	2	3	4	5	6	7	8	9
-3.0	0.0013	0.0010	0.0007	0.0005	0.0003	0.0002	0.0002	0.0001	0.0001	0.0000
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0026	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0084
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0119
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0238	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0300	0.0294
-1.7	0.0446	0.0436	0.0426	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0085	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0570	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.01711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2909	0.2061	0.2033	0.2266	0.1977	0.1949	0.1922	0.1894	0.1857
-0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2005	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2276
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3281	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

* B.W. Lindgren, **Statistical Theory.** The Macmillan Company, 1960.

(Contd.)

Table 1. (Contd.)

$$\phi(z) = \int_{-x}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = P(Z \leq z)$$

z	0	1	2	3	4	5	6	7	8	9
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	0.55910	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	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	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	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9317
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	0.9874	0.9878	0.9881	0.9984	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	0.9956	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	0.9871	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

UNIT 12 BIVARIATE DISTRIBUTIONS

Structure

12.1 INTRODUCTION

In the Units 10 and 11, we have discussed some univariate continuous distributions. These are examples of probability distributions of one-dimensional random variables. There are situations when it is important to study more than one characteristic at the same time. For instance, in medical studies, the health of a patient might depend on the nutrition intake, the cholesterol level etc. In meterological data, the amount of rainfall in a period might depend on temperature, atmospheric pressure etc. In marketing a common product, the sales might depend on the price of the product, competition from similar products, purchasing power of people to which the product is directed etc. In other words, simultaneous study of characteristics is needed at times instead of their study individually to observe the interrelations between different characteristics. We restrict to the case of two characteristics in this unit.

In Section 12.2, we introduce the notion of bivariate distribution and present some examples. The concept of marginal distributions for individual components is discussed in this section. Conditional distributions of one variable given the other, plays a major role in the study of bivariate distributions. These distributions describe the probabilistic behaviour of one variable when the other variable is fixed. If there is no change in the probabilistic behaviour of one given the other, then the components are said to be independent. Sections 12.3 and 12.4 contain discussion on conditional distributions and independence. There are many concepts associated with bivariate random variables which give information about the relationship between them or about the location, symmetry of the joint density function. Some of the more important of these concepts are expectation, moment, covariance, correlation and regression. In Sections 12.5 and 12.6 we introduce you to these concepts.

Objectives

After reading this unit, you should be able to :

- compute marginal and conditional distributions given a bivariate distribution;
- compute expectation of a function of a random vector, moments, moment generating functions, covariance correlation coefficient and regression of one component given the other.

12.2 THE DENSITY FUNCTION OF A BIVARIATE DISTRIBUTION

Let us start with the following situation :

Suppose we want information about the life time of a machine in a factory. Suppose a machine in a factory has just two components. Let X and Y denote the life times of

these components. Let us further assume that the machine works provided both the components function properly. In other words the life time of the machine (say) Z depends on the life times of the components X and Y. Here the knowledge of the probability distributions of X and Y separately is not enough to get information on Z. It needs to be known whether the performance of the first component affects the performance of the second component and vice versa, that is, the probability distribution of Z depends jointly on the distribution of X and Y.

There are other situations where joint behaviour of two or more components affects the performance or behaviour of the whole system. For instance, in medical studies, say, of persons suffering from heart diseases, factors like blood pressure, cholesterol etc. play important role in the status of their health. Here also the components individually will not give a true picture.

In all the above situations, you must have noted that there are several instances where one needs to study the interrelation between random variables. This leads us to the notion of multivariate distributions. For simplicity, we consider the case of bivariate continuous distributions in detail. Recall that you have studied some bivariate distributions in the discrete case in Unit 7, Block 2.

Now we shall define a bivariate random variable and its distribution function.

Definition : Suppose X and Y are two continuous r.v.'s defined on a sample space Ω. Then the function (X, Y) defined on Ω by

$$(X, Y)(\omega) = (X(\omega), Y(\omega))$$

is called the random vector or a bivariate random variable.

The joint probability distribution of this r.v., denoted by $F_{X, Y}$ is defined as

$$F_{X, Y}(x, y) = P[X \leq x, Y \leq y].$$

Note that $[X \leq x, Y \leq y]$ stands for the event $[X \leq x] \cap [Y \leq y]$.

Now we shall define the density function of a joint distribution. From the univariate case you recall that a distribution function is said to be absolutely continuous if there exists a non-negative real-valued function f such that

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

and the function f is called the density function of F. In the same way a joint distribution is said to be absolutely continuous if there exists a non-negative function $f_{X, Y}(x, y)$ such that

$$F_{X, Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X, Y}(u, v)\, du\, dv, \qquad -\infty < x, y < \infty.$$

The function $f_{X, Y}(x, y)$ is called the **joint probability density function** or joint density in short of (X, Y).

Let us first take a look at some elementary properties of the joint distribution

Recall that you have seen simultaneous limits $\lim_{\substack{x \to a \\ y \to b}} f(x, y)$ of a function of two variables in Unit 4 of MTE–07.

i) $F_{X, Y}(-\infty, -\infty) = \lim_{\substack{x \to -\infty \\ y \to -\infty}} F_{X, Y}(x, y) = 0$

ii) $F_{X, Y}(\infty, \infty) = \lim_{\substack{x \to \infty \\ y \to \infty}} F_{X, Y}(x, y) = 1$

Now let us denote by $F_X(x)$ and $F_Y(y)$ the distribution functions of X and Y respectively. Is there any relationship between F_X, F_Y and $F_{X,Y}$? Let's see.

From the definition of $F_X(x)$

$$\begin{aligned} F_X(x) &= P[X \le x] \\ &= P[X \le x, Y < \infty] \\ &= \lim_{y \to \infty} P[X \le x, Y \le y] = \lim_{y \to \infty} F_{X,Y}(x, y). \end{aligned}$$

The last statement is intuitively clear as the set $[X \le x, Y \le y]$ tends to the set $[X \le x, Y < \infty]$ as $y \to \infty$. However a rigorous argument can be given by using the result, if $A_n \to A$, then $P(A_n) \to P(A)$ as $n \to \infty$. We will not go into the discussion here. Let us denote the limit of $F_{X,Y}(x, y)$ as $y \to \infty$ by $F_{X,Y}(x, \infty)$. Hence

$$F_X(x) = F_{X,Y}(x, \infty).$$

Similarly

$$F_Y(y) = F_{X,Y}(\infty, y).$$

The distribution functions $F_X(x)$ and $F_Y(y)$ are called the **marginal distribution** of X and Y respectively.

Then with the help of the marginal distribution we can easily compute the probability that X is greater than x and Y is greater than y. From elementary theorems of probability (see Unit 5), it follows that

$$\begin{aligned} P[X > x, Y > y] &= 1 - P\left[[X > x, Y > y]^C\right] \\ &= 1 - P\left[[X > x]^C \cup [Y > y]^C\right] \\ &= 1 - P[[X \le x] \cup [Y \le y]] \\ &= 1 - [P[X \le x] + P[Y \le y] - P[X \le x, Y \le y]] \\ &= 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y). \end{aligned}$$

Here we have used the fact

$$(A \cap B)^C = A^C \cup B^C$$

for any two sets A and B and the results

$$\begin{aligned} P(A^C) &= 1 - P(A), \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B), \end{aligned}$$

from the elementary theorem of probability in Unit 5. You should check that, for any $a_1 < a_2$ and $b_1 < b_2$,

$$\begin{aligned} &P[a_1 < X \le a_2, b_1 < Y \le b_2] \\ &\quad = F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1). \end{aligned}$$

Let us now suppose that (X, Y) is a bivariate random vector with density function $f_{X,Y}(x, y)$. From the properties of joint distribution function, we have

i) $f_{X,Y}(x, y) \ge 0, \quad -\infty < x, y < \infty,$

ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1$, and

iii) $P[X \le x, Y \le y]] = P[(X, Y) \in (-\infty, x] \times (-\infty, y]]$

$$\begin{aligned} &= F_{X,Y}(x, y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) du\, dv. \end{aligned}$$

In fact it can be shown that Property (iii) holds not only for sets of the form $[a, b] \times [c, d]$ but for more general sets. We have the result, for any set C in $\mathbf{R}^2$, the two dimensional plane,

$$P[(X, Y) \in C] = \int\int_{(u, v) \in C} f_{X, Y}(u, v)du\, dv \qquad \text{.....(1)}$$

(The set C is not as general as we have stated. Some technical restrictions have to be made.)

Now recall that in the univariate case we have shown that if F(x) and f(x) are the distribution function and density function of a r.v. X, then

$$\frac{d}{dx} F(x) = f(x)$$

(see Sec 10.3 of Unit 10).

You may ask here, what is the analogue of this in the bivariate case ? In the bivariate case we have the following result.

Suppose that (X, Y) is a bivariate random vector with density function $f_{X, Y}(x, y)$ and the distribution function $F_{X, Y}(x, y)$. Then $F_{X, Y}(x, y)$ has second order partial derivatives with respect to x and y for almost all (x, y) in the plane and

$$\frac{\partial^2 F_{X, Y}(x, y)}{\partial x\, \partial y} = f_{X, Y}(x, y).$$

This result follows from the properties of double integrals.

Next we shall discuss how we can derive the density functions of X and Y individually from the joint density function $f_{X, Y}$. For that let us consider the marginal distributions of X and Y. Let f_X and f_Y denote the density functions of X and Y respectively. Then we have

$$\frac{d}{dx} F_X(x) = f_X(x) \qquad \text{.....(2)}$$

for almost all x,

$$\frac{d}{dx} F_Y(y) = f_Y(y) \qquad \text{.....(3)}$$

for almost all y.

Also we have

$$\begin{aligned} F_X(x) &= F_{X, Y}(x, +\infty) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X, Y}(u, v)du\, dv \\ &= \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f_{X, Y}(u, v)dv\right] du \\ &= \int_{-\infty}^{x} f_X(u)du \end{aligned}$$

where $$f_X(u) = \int_{-\infty}^{\infty} f_{X, Y}(u, v)dv.$$

Then from (2) we have

$$f_X(x) = \int_{-\infty}^{\infty} f_{X, Y}(x, y)\, dy \quad -\infty < x < \infty \qquad \text{.....(4)}$$

Similarly

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx, \quad -\infty < y < \infty \qquad .....(5)$$

This shows that if (X, Y) has a joint density function $f_{X,Y}(x, y)$, then X has a probability density function given by (4) and Y has probability density function given by (5). The functions $f_X(x)$ and $f_Y(y)$ are called the **marginal densities** of X and Y respectively.

Let's see some examples now.

Example 1 : Suppose (X, Y) has joint density function

$$f(x, y) = \begin{cases} 2\,e^{-x}e^{-2y} & , \ 0 < x, y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

Let us compute the marginal density $f_X(x)$ and compute $P[X < y]$.

By the discussion we have earlier,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

$$= 2e^{-x} \int_0^{\infty} e^{-2y}\,dy$$

$$= 2e^{-x} \left[\frac{1}{2} e^{-2y}\right]_0^{\infty}$$

$$= e^{-x} \qquad \text{for } x > 0$$

and $\qquad f_X(x) = 0 \qquad \text{for } x \le 0$

Now let us compute $P[X < y]$. For that we first describe the event $[X < y]$ in the plane. Let $C = [X < y]$. Note that we can write C as $[0 < X < y, 0 < Y < \infty]$. Then by (1)

$$P[X < y] = \iint_C f_{X,Y}(u, v)du\,dv$$

$$= \int_0^y \int_0^{\infty} f_{X,Y}(x, y)dx\,dy$$

$$= \int_0^{\infty} \left\{ \int_0^y 2\,e^{-x}\,e^{-2y}\,dx \right\} dy$$

$$= \int_0^{\infty} 2\,e^{-2y} \left[-e^{-x}\right]_0^y dy$$

$$= \int_0^{\infty} 2\,e^{-2y} \left[1 - e^{-y}\right]_0^y dy$$

$$= \int_0^{\infty} 2\,e^{-2y} - \int_0^{\infty} 2\,e^{-3y}\,dy$$

$$= \left[-e^{-2y}\right]_0^{\infty} - \left[-\frac{2}{3} e^{-3y}\right]_0^{\infty}$$

$$= 1 - \frac{2}{3} = \frac{1}{3}.$$

Example 2 : Suppose (X, Y) has the uniform density over the unit circle of radius r with center at (0, 0). In other words, (X, Y) has the joint density function

$$f(x, y) = C \quad \text{if } 0 \le x^2 + y^2 \le r^2$$
$$= 0 \qquad \text{otherwise.}$$

where C is a constant. Let us compute the marginal densities $f_X(x)$ and $f_Y(y)$.

Since f is the joint density function we have $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dx\,dy = 1$. This implies that

$$C \iint_D dx\,dy = 1$$

where $D = \{(x, y) : x^2 + y^2 \le r^2\}$

Now, what is this double integral represent ? From your knowledge of double integrals (see MTE–07, Block 4) you know that this double integral represents the surface area of D. That means the double integral represents the area of a circle of radius r with center (0, 0) which is πr^2. Hence

$$C = \frac{1}{\pi r^2}.$$

Let us now compute the marginal density of X. Here

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

$$= \frac{1}{\pi r^2} \int_{\{(x, y) : x^2 + y^2 \le r^2\}} dy$$

$$= \frac{1}{\pi r^2} \int_{-\sqrt{r^2 - x^2}}^{+\sqrt{r^2 - x^2}} dy$$

$$= \frac{2}{\pi r^2}\sqrt{r^2 - x^2} \qquad \text{for } 0 \le x^2 \le r^2$$

and

$$f_X(x) = 0 \qquad \text{otherwise.}$$

Similarly you can get that

$$f_Y(y) = \frac{2}{\pi r^2}\sqrt{r^2 - y^2} \qquad \text{for } 0 \le y^2 \le r^2$$

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

In the next example we introduce you to an important bivariate distribution known as bivariate normal distribution.

Example 3 : A random vector (X, Y) is said to follow a bivariate normal distribution if it has the joint density given by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} \exp\left(-\frac{q}{2}\right), \quad -\infty < x, y < \infty.$$

where σ_x and σ_y assume values in the interval $]0, \infty[$, ρ assumes values in the interval $]-1, 1[$, and

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right] \quad \text{.....(6)}$$

Let us find the marginal densities of X and Y.

By definition the marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{q}{2}\right)$$

where q is as in (6). We first rewrite the expression (6) in the following form

$$q = \frac{1}{\sigma_y^2(1-\rho^2)}\left[y - \left\{\mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)\right\}\right]^2 + \frac{(x-\mu_x)^2}{\sigma_x^2}$$

Then we get

$$f_X(x) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}}$$

$$\times \int_{-\infty}^{\infty} \exp\left\{\frac{1}{2\sigma_y^2(1-\rho^2)}\left[y - \left\{\mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)\right\}\right]^2\right\} dy$$

$$= \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}}$$

$$\times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_y^2(1-\rho^2)}} \exp\left(-\frac{1}{2\sigma_y^2(1-\rho^2)}\left[y - \left\{\mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)\right\}\right]^2\right) dy$$

The integrand in the integral on the right side can be seen to be normal density function with mean

$$\mu_y + \frac{\rho\sigma_y}{\sigma_x}(x-\mu_x)$$

and variance $\sigma_y^2(1-\rho^2)$ for any fixed x. Hence it integrates to unity and therefore

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}$$

Thus the marginal density of X is normal with mean μ_x and variance σ_x^2. Similarly you can show that the marginal density of Y is normal with mean μ_y and variance σ_y^2. We leave it as an exercise for you (see E6).

Why don't you try some exercises now.

E1) In a statistical survey, it was found that if X denotes the daily number of hours a child watches television and Y denotes the number of hours he or she spends on the studies, then (X, Y) has the joint density function

$$f(x, y) = xy\, e^{-(x+y)} \quad \text{if } x > 0,\ y > 0$$
$$= 0 \quad \text{otherwise.}$$

What is the probability that a child chosen at random spends at least twice as much time watching television as he or she does on studies ?

E2) Consider an electronic system with two components. Suppose the system is such that one component is on the reserve and it is activated only when the other component fails. Let X and Y denote the life times of these components. Suppose the system fails if and only if both the components fail. (X, Y) has the joint density function

$$f(x, y) = \lambda^2 e^{-\lambda(x+y)} \quad \text{if } x \geq 0, y \geq 0$$
$$= 0 \quad \text{otherwise.}$$

What is the probability that the system will last for more than 500 hours ?

E3) Find the joint density function of (X, Y) when its joint distribution function is given by

$$F_{X,Y}(X, Y) = (1 - e^{-\lambda y})(1 - e^{-\lambda x}) \quad \text{if } x > 0, y > 0$$
$$= 0 \quad \text{otherwise.}$$

E4) Suppose (X, Y) has the joint density function

$$f(x, y) = \begin{cases} y^2 e^{-y(x+1)} & \text{if } x \geq 0, y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Determine the marginal densities $f_X(x)$ and $f_Y(y)$ of X and Y respectively.

E5) Compute the marginal density of Y, for the bivariate normal distribution.

So far in this section we have discussed the distribution functions and joint density functions of bivariate random variables. These functions allow us to obtain answers to probabilistic questions pertaining to the random variables. In the next section we shall discuss another distribution function related to a random vector.

12.3 CONDITIONAL DISTRIBUTIONS

Let (X, Y) be a bivariate random vector. There are situations when we would like to know whether a change in the value of Y has influence on X. In other words, for instance, if Y increases, does X also increase or if Y increases, does X decrease etc. ?

For example, suppose X denotes the height of a person and Y his or her weight. A natural question is to examine the relation between the height and the weight. In such cases, it is important to know the behaviour of X for a given value of Y and vice versa. In order to study problems of this nature we introduce the concept of conditional distributions. Before that let us recall the definition of the conditional probability in the case of bivariate discrete random variables from Unit 7.

The conditional probability of the event [Y = y] given the event [X = x] is defined as

$$P[Y = y \mid X = x] = \frac{P[X = x, Y = y]}{P[X = x]} \qquad \text{.....(7)}$$

provided $P[X = x] \neq 0$.

Can we directly use this definition for continuous random variables ? Obviously we can't, because you know that in this case P[X = x] = 0 for any x. Now, suppose we rewrite (7) in the following way :

$$P[Y = y \mid X = x] = \frac{f_{X,Y}(x, y)}{f_X(x)} \qquad \text{.....(8)}$$

where $f_{X,Y}(x, y)$ denotes the joint probability mass function and $f_X(x)$ denotes the marginal probability mass function of X [see Unit 7, Block 2]. If we replace $f_{X,Y}(x,y)$ and f(x) in (7) by the joint density function of the random vector (X, Y) and the marginal density function of X respectively in the continuous case, then we can define the conditional probability analogous to (8).

We make the following definition :

Definition : Let (X, Y) be a bivariate random vector with density function $f_{X, Y}(x, y)$. Let $f_X(x)$ and $f_Y(y)$ denote the marginal density functions of X and Y respectively. Then the conditional density function of X given Y = y which we denote by $f_{X|Y}(x|y)$ defined as

$$f_{X|Y}(x \mid y) = \frac{f_{X, Y}(x, y)}{f_Y(y)} \quad , -\infty < x < \infty$$

for any y such that $f_Y(y) > 0$.

Similarly the conditional density function of Y given X = x is defined as

$$f_{Y|X}(y \mid x) = \frac{f_{X, Y}(x, y)}{f_X(x)} \quad , -\infty < y < \infty$$

for any x such that $f_X(x) > 0$.

We would like to remind you again that the functions $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ are well-defined whenever $f_Y(y) > 0$ and $f_X(x) > 0$ respectively even though P[X = x] and P[Y = y] is zero for all x and y.

Next we shall see whether the conditional densities of X given Y = y and of Y given X = x are genuine probability density functions i.e. to check whether they satisfy the conditions

i) $f_{X|Y}(x|y) \geq 0$, $f_{Y|X}(y|x) \geq 0$.

ii) $\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)dx = 1$ and $\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)dy = 1$.

For instance, let us check conditions (i),(ii) for $f_{X|Y}(x|y)$. From the definition of $f_{X|Y}(x|y)$ we get that $f_{X|Y}(x|y)$ is non-negative for all x. Also

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} \frac{f_{X, Y}(x, y)}{f_Y(y)} dx$$

$$= \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X, Y}(x, y)\, dx$$

$$= \frac{f_Y(y)}{f_Y(y)} = 1$$

for any y such that $f_Y(y) > 0$.

Similarly you can show that $f_{Y|X}(y \mid x) \geq 0$ and

$$\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)dy = 1$$

for every x with $f_x(x) > 0$.

Now that we have defined the conditional density functions, we can easily define the conditional distribution functions as follows :

The **conditional distribution function** of X given Y = y is defined by

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(u \mid y)\, du, \qquad -\infty < x < \infty$$

and the conditional distribution function of Y given X = x is defined by

$$F_{Y|X}(y \mid x) = \int_{-\infty}^{y} f_{Y|X}(v \mid x)\, dv, \quad -\infty \le y \le \infty$$

Let us consider some examples now.

Example 4 : Suppose the random vector (X,Y) has the joint density function

$$f_{X,Y}(x, y) = \begin{cases} 2 & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Let us now compute the marginal densities and conditional densities. Then determine the probability that $P\left[0 < X < \frac{1}{2} \;\middle|\; Y = \frac{3}{4}\right]$.

The marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$$

$$= \int_{x}^{1} 2\, dy \quad = 2(1 - x) \qquad \text{if } 0 < x < 1$$

and

$$f_X(x) = 0 \qquad \text{otherwise.}$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$$

$$= \int_{0}^{y} 2\, dx = 2y \qquad \text{for } 0 < y < 1$$

and

$$f_Y(y) = 0 \qquad \text{otherwise.}$$

The conditional density of X given Y = y is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

$$= \frac{2}{2y} \qquad \text{if } 0 < x < y$$

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

whenever $0 < y < 1$.

Similarly the conditional density of Y given X = x is

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

$$= \frac{2}{2(1 - x)} \qquad \text{if } x < y < 1$$

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

whenever $0 < x < 1$.

Let us now compute $P\left[0 < X < \frac{1}{2} \mid Y = \frac{3}{4}\right]$ using the marginal densities. By definition

$$P\left[0 < X < \frac{1}{2} \mid Y = \frac{3}{4}\right] = \int_0^{1/2} f_{X|Y}\left(x \mid y = \frac{3}{4}\right) dx$$

Note that for $y = 3/4$, $f_{X|Y}(x \mid y) = \frac{2}{2 \times 3/4} = \frac{4}{3}$. Therefore

$$P\left[0 < X < \frac{1}{2} \mid Y = \frac{3}{4}\right] = \int_0^{1/2} \frac{4}{3} dx = \frac{2}{3}.$$

However

$$P[0 < X < 1/2] = \int_{-\infty}^{\infty} f_X(x)\, dx$$

$$= \int_0^{1/2} 2(1 - x)\, dx = 3/4.$$

Hence the conditional probability that X lies between 0 and 1/2 given Y = 3/4 is different from the unconditional probability that X lies between 0 and 1/2.

Example 5 : Suppose the random vector (X, Y) follows bivariate normal distribution. Let us obtain the conditional distribution of X given Y = y and that of Y given X = x.

By definition the conditional density of Y given X=x is

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

$$= \frac{1}{\sqrt{2\pi\, \sigma_y^2 (1 - \rho^2)}}$$

$$\times \exp\left(\frac{1}{2\, \sigma_y^2 (1 - \rho^2)}\left[y - \left\{\mu_y + \frac{\rho\, \sigma_y (x - \mu_x)}{\sigma_x}\right\}\right]^2\right) \quad \ldots (9)$$

You might have noticed that this function also looks like a normal density function. Infact it is just the normal density function with mean

$$\mu_y + \frac{(\rho\sigma_y)}{\sigma_x}(x - \mu_x)$$

and variance

$$\sigma_y^2 (1 - \rho^2).$$

In other words, the conditional density of Y given X = x is normal with mean $\mu_y + \frac{\rho\sigma_y}{\sigma_x}(x - \mu_x)$ and variance $\sigma_y^2 (1 - \rho^2)$. Similarly the conditional density of X given Y = y is normal with mean $\mu_x + \frac{\sigma_x}{\sigma_y}(y - \mu_y)$ and variance $\sigma_x^2 (1 - \rho^2)$.

To get more practise why don't you do some exercises.

E6) Suppose that the joint density function of (X, Y) is

$$f(x, y) = \begin{cases} C(x + y^2) & , \quad \text{for } 0 < x, y < 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Find the conditional probability density function of X given Y = y for $0 < y < 1$.

Then compute $P\left(X < \frac{1}{2} \mid Y = \frac{1}{2}\right)$.

E7) Suppose that the joint density function of (X,Y) is

$$f(x, y) = 2e^{-(x+y)} \qquad , 0 < x < y$$
$$= 0 \qquad \text{otherwise.}$$

Compute $P[Y < 1 \mid X < 1]$ and $P[Y < 1 \mid X = 1]$.

E8) Suppose that the conditional density function of Y given x = x and the marginal density function of X are given by

$$f_{Y|X}(y \mid x) = \frac{2y + 4x}{1 + 4x} \qquad , 0 < x, y < 1$$
$$= 0 \qquad \text{otherwise}$$

and

$$f_X(x) = \frac{1 + 4x}{3} \qquad , 0 < x < 1$$
$$= 0 \qquad \text{otherwise}$$

respectively. Determine the marginal density function of Y.

E9) Let X denote the percentage of marks obtained by a student in Mathematics in Class XII final examination and Y denotes the marks obtained in English. Suppose that (X,Y) has the joint density function

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y) & \text{if } 0 < x, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

What percentage of students obtain more than 80% in Mathematics ? If a student has obtained 30% in English, what is the probability that he gets more than 80% in Mathematics ? If a student has obtained 30% in Mathematics, what is the probability that he gets more than 80% in English ?

In the next section we shall define "Independence" of two continuous random variables by analogy with the discrete case (see Unit 7, Block 2).

12.4 INDEPENDENCE

Suppose (X,Y) is a bivariate random vector with the joint density function $f_{X,Y}(x, y)$ and the marginal densities $f_X(x)$ and $f_Y(y)$. Let us denote the conditional densities of X given Y = y by $f_{X|Y}(x \mid y)$ and of Y given X = x by $f_{Y|X}(y \mid x)$ as before.

Let us now suppose that $f_{X|Y}(x \mid y)$ does not depend on y. Then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$$

$$= \int_{-\infty}^{\infty} \frac{f_{X,Y}(x, y)}{f_Y(y)} f_Y(y)\, dy \quad , \text{ for y with } f_Y(y) > 0$$

$$= \int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\, f_Y(y)\, dy$$

$$= f_{X|Y}(x \mid y) \int_{-\infty}^{\infty} f_Y(y) dy = f_{X|Y}(x \mid y)$$

since $f_{X|Y}(x \mid y)$ is independent of y. Hence

$$f_X(x) = f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \text{ for y with } f_Y(y) > 0 \qquad \ldots(10)$$

In other words, the marginal density of X and the conditional density of X given Y=y are the same for every y with $f_Y(y) > 0$. In particular, it follows that the conditional distribution of X given Y = y does not depend on y since

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(u \mid y)\, du$$

$$= \int_{-\infty}^{x} f_X(u)\, du = F_X(x)$$

These observations in fact tell us that the probabilistic behaviour of X does not depend on the value of Y that is observed.

Let us again look at Equation (10). This equation can also be written in the form

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y), \qquad -\infty < x, y < \infty$$

which is symmetric in X and Y. Such pairs of random variables are called independent random variables. More precisely we have the following definition.

Definition : Suppose (X,Y) is a bivariate random vector with the joint density function $F_{X,Y}(x, y)$ and the marginal densities $f_X(x)$ and $f_Y(y)$ respectively. The random variables X and Y are said to be **independent** if

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y), \qquad -\infty < x, y < \infty \qquad ...(11)$$

Let us now consider two independent random variables X, Y. Let us denote by $F_{X,Y}(x, y)$, $F_X(x)$ and $F_Y(y)$ the joint distribution of (X, Y), the marginal distribution of X and the marginal distribution of Y respectively. From the definition of $F_{X,Y}(x, y)$ and $f_{X,Y}(x, y)$, you note that

$$F_{X,Y}(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)\, du\, dv$$

$$= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)\, f_Y(v)\, du\, dv$$

$$= \left\{\int_{-\infty}^{x} f_X(u)\, du\right\}\left\{\int_{-\infty}^{y} f_Y(v)\, dv\right\}$$

$$= F_X(x)\, F_Y(y), \quad -\infty < x, y < \infty.$$

Hence, if X and Y are independent random variables, then

$$F_{X,Y}(x, y) = F_X(x)\, F_Y(y), \qquad -\infty < x, y < \infty. \qquad ...(12)$$

In other words,

$$P[X \le x, Y \le y] = P[X \le x]\, P[Y \le y]$$

for all x and y. This relation shows that the events $[X \le x]$ and $[Y \le y]$ are independent events (Recall the definition of independence of events from Unit 5 Block 2) for all x and y if X and Y are independent. In fact the converse is also true. That means if the events $[X \le x]$ and $[Y \le y]$ are independent for all x and y, then X and Y are independent provided $F_{X,Y}$ is absolutely continuous. Why don't you check this for yourselves ? (see E10).

The above discussion indicates that the definition of independent random variables is consistent with the definition of independent events. In fact there may be some advantage in defining the independence of r.v.'s X and Y using the independence of events $[X \le x]$ and $[Y \le y]$ for every pair (x,y). This definition is more general and does not require the existence of the joint density function.

Let us now look at the independence of certain random vectors we have already come across. For example, consider those in Example 4 and E3. Are they independent ? You can easily see that the random variables in Example 4 are not independent whereas the random variables in E3 are independent.

Let us consider another example.

Example 6 : Suppose (X, Y) is a bivariate random vector with density function

$$f_{X,Y}(x, y) = \frac{1}{2}\pi e^{-\left(\frac{x^2+y^2}{2}\right)}, \quad -\infty < x, y < \infty.$$

Let us check whether X and Y are independent.

We first calculate the marginal density function $f_X(x)$ and $f_Y(y)$. We have

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\left(\frac{x^2+y^2}{2}\right)} dy$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad , -\infty < x < \infty$$

since

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = 1$$

being the integral of standard normal density function. Similarly, we have

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \qquad , -\infty < y < \infty.$$

Hence

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y), \qquad -\infty < x, y < \infty$$

which proves that X and Y are independent random variables.

Here is another example.

Example 7 : Suppose (X,Y) is a bivariate random vector with joint density function

$$f_{X,Y}(x, y) = 8xy, \quad 0 < x < y < 1$$
$$= 0 \quad \text{otherwise}$$

Let us check whether X and Y are independent.

Then

$$f_X(x) = \int_x^1 8xy\, dy \qquad \text{for } 0 < x < 1$$

$$= 8x \left[\frac{y^2}{2}\right]_x^1$$

$$= 4x(1 - x^2) \qquad \text{for } 0 < x < 1$$

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

Similarly

$$f_Y(y) = \int_0^y 8xy\,dx \qquad \text{for } 0 < y < 1$$

$$= 8y\left\{\frac{x^2}{2}\right\}_0^y \qquad \text{for } 0 < y < 1$$

$$= 4y^3 \qquad \text{for } 0 < y < 1$$

$$= 0 \qquad \text{otherwise.}$$

Obviously

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

is not satisfied for all x,y. Hence X and Y are not independent random variables.

This example indicates that random variables X and Y are dependent even though the joint density function $f_{X,Y}(x, y)$ is the product of a function of x and a function of y. You should note that the set $[(x, y) : 0 < x < y < 1]$ where the joint density is positive is not a product set, that is a set of the form $A_1 \times A_2$ where A_1 is a set defined by x alone and A_2 is a set defined by Y alone. You can try some exercises now.

E10) Prove that if the events $[X \leq x]$ and $[Y \leq y]$ are independent for all x and y, then X and Y are independent.

E11) Suppose (X, Y) is a random vector with the joint density function

$$f(x, y) = \begin{cases} 12xy(1-y) & 0 < x, y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

Show that X and Y are independent random variables.

E12) Suppose (X,Y) has the joint density function

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

Determine $P[0 < X < 1/3, 0 < Y < 1/3]$

E13) Suppose the joint density of (X,Y) is given by

$$f(x, y) = \begin{cases} x\, e^{-(x+y)} & x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Are X and Y independent ?

Here we shall make a remark.

Remark : If X and Y are independent random variables, then it can be shown that g(X) and h(Y) are independent random variables where g(X) and h(Y) are any functions of the random variables X, Y respectively. We omit the proof. There are technical restrictions on g and h but these conditions are generally satisfied by the function which we come across in this course. We will not discuss its proof as it is beyond the scope of this course. If you are interested you can find the proof in the reference books.

The exercises in this section and in the earlier sections would have given you enough practice to compute the density functions and distribution functions of bivariate random variables. Next we shall discuss some measures of central tendancy of the probability distribution function of bivariate random vectors.

12.5 EXPECTATIONS AND MOMENTS OF FUNCTIONS OF A RANDOM VECTOR

Let us consider a bivariate random vector (X,Y) with density function $f_{X,Y}(x, y)$. How do we define the expectation of a bivariate random vector ? Do you think that it

is E(XY) or is it E(X + Y) ? This leads us to define the expectation of any function g(X, Y) of X and Y. In this unit g(X, Y) means only simple functions like X + Y, XY, |X – Y| etc. In the next unit we shall talk about the functions g(X, Y) of a bivariate random vector in detail.

We shall begin with the definition of the expectation of a function of a r.v.

Definition : Let g(X,Y) be a function of bivariate random vector with joint density function $f_{X, Y}(x, y)$. Then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)\, f_{X, Y}(x, y)\, dx\, dy$$

whenever

$$E[|g(X, Y)|] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g(x, y)|\, f_{X,Y}(x, y)\, dx\, dy$$

is finite.

Now let us calculate the expectation of g(X,Y) = X+Y where X and Y are two random variables such that E(X) and E(Y) are finite.

Suppose g (X,Y) = X+Y and E(X), E(Y) are finite. Then

$$E(X + Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)\, f_{X, Y}(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{X, Y}(x, y)\, dx \right] dx + \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} f_{X, Y}(x, y)\, dx \right] dy$$

$$= \int_{-\infty}^{\infty} x\, f_X(x)\, dx + \int_{-\infty}^{\infty} y\, f_Y(Y)\, dy$$

$$= E(X) + E(Y).$$

All the above calculations can be justified under the assumption that E(X) and E(Y) are finite. Hence we have the following important property of expectations.

Theorem 1 : If X and Y are random variables such that E (X) and E(Y) are finite, then E(X+Y) is finite and

$$E(X+Y) = E(X) + E(Y).$$

This property can be extended to any finite number of random variable $X_1, X_2, \ldots, X_n$ by using mathematical induction.

Let us now suppose that

$$g(X, Y) = (X - a)^r (Y - b)^s$$

where a and b are some constants. Suppose that E (| g(X, Y) |) is finite. Then

$$E[g(X, Y)] = E\left[(X - a)^r (Y - b)^s\right]$$

is called the **product moment** of (X, Y) about (a, b) of order (r, s). Let us choose $a = \mu_X = E(X)$ and $b = \mu_Y = E(Y)$ and r, s to be integers greater than or equal to zero. If r = 2 and s = 0, then the corresponding moment is Var (X). If r = 0 and s = 2, then the corresponding moment is Var (Y).

If r = 1 and s = 1 then

$$E[g(X, Y)] = E[(X - \mu_X)(Y - \mu_Y)]$$

is called the **covariance** of X and Y and it is denoted by Cov (X, Y). You recall that you are already familiar with the covariance of two discrete random variables from Unit 7 Block 2. There we have shown that

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y).$$

Does this expression hold for continuous case also ? Why don't you check it for yourself in the following exercise.

E14) Show that $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$

If you have done E15, you must also have noticed that the covariance between X and Y is the same as the covariance between Y and X. In other words covariance is symmetric in X and Y. It is sometimes convenient to observe that Cov (X,X) = Var (X). In general the covariance between X and Y is a measure of the relationship between them. If X tends to be large when Y is large and tends to be small when Y is small, then the covariance is positive. On the other hand if large values of X correspond to small values of Y or small values of X correspond to large values of Y, then the covariance is negative.

Let us see some examples now.

Example 8 : Suppose (X,Y) has the joint probability density function

$$f_{X,Y}(x, y) = \begin{cases} 8xy & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

with $E(X) = \frac{8}{15}$, $E(Y) = \frac{4}{5}$. Let us compute the covariance of X and Y.

By definition

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

$$E(XY) = \int_0^1 \int_0^y xy\, 8xy\, dx\, dy$$

$$= \int_0^1 8y^2 \left\{ \int_0^y x^2\, dx \right\} dy$$

$$= \int_0^1 8y^2 \frac{y^3}{3}\, dy = \frac{8}{3}\left[\frac{y^6}{6}\right]_0^1$$

$$= \frac{4}{9}$$

Hence

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

$$= \frac{4}{9} - \frac{8}{15} \times \frac{4}{5}$$

Here we ask you a question : What happens to the covariance of X and Y when X and Y are independent ? You can find an answer to this question in the following theorem.

Theorem 2 : If X and Y are independent random variables with E(X) and E(Y) finite, then E(XY) exists and

$$E(XY) = E(X)\,E(Y) \qquad \ldots(13)$$

Proof : Suppose X and Y are independent random variables with the density functions $f_X(x)$ and $f_Y(y)$ respectively. From the definition of independence, it follows that the joint density of (X, Y) is given by

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y).$$

Suppose that E(X) and E(Y) exist. Then

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x, y)\, dxy\, dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x)\, f_Y(y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} x\, f_X(x) \left[\int_{-\infty}^{\infty} y\, f_Y(y)\, dy\right] dx$$

$$= E(Y)\int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

$$= E(Y)\, E(X).$$

From the theorem we get that if X and Y are independent with finite expectation, then

$$Cov(X, Y) = E(XY) - E(X)\, E(Y) = 0$$

An immediate consequence of the multiplication law for expectation is the following result concerning the variance of sum of two independent random variables.

Theorem 3 : If X and Y are independent random variables with finite variances, then $Var(X + Y) = Var\, X + Var\, Y$

Proof : Let $E(X) = \mu_X$ and $E(Y) = \mu_Y$. Then

$$E(X + Y) = \mu_X + \mu_Y$$

and

$$Var(X + Y) = E\left[(X + Y - (\mu_X + \mu_Y))^2\right]$$

$$= E\left[(X - \mu_X) + (Y - \mu_Y)^2\right]$$

$$= E\left[X - \mu_X\right]^2 + E\left[Y - \mu_Y\right]^2 + 2\, E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

$$= Var(X) + Var(Y) + 2\, Cov(X, Y).$$

Since the random variables X and Y are independent, it follows that Cov (X,Y) = 0 and hence

$$Var(X + Y) = Var(X) + Var(Y).$$

From the proof given above, you must have noticed that the above result concerning the variance of sum of two random variables is true if Cov (X,Y) = 0.

Let us consider another example.

Example 9 : Suppose (X,Y) has joint density function

$$f_{X,Y}(x, y) = \begin{cases} 1 & \text{if } -y < x < y,\ 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let us first compute the marginal densities $f_X(x)$ and $f_Y(y)$ respectively and the covariance of X and Y.

We have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$$

$$= \int_{-y}^{y} dx$$

$$= 2y \qquad \text{for } 0 < y < 1$$

$$= 0 \qquad \text{elsewhere}$$

and

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

$$= \int_{-x}^{1} dy \qquad \text{if } -1 < x < 0$$

Then

$$f_X(x) = \int_{-x}^{1} dy\,, \qquad \text{if } -1 < x < 0$$

$$= \int_{x}^{1} dy\,, \qquad \text{if } 0 < x < 1$$

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

That is,

$$f_X(x) = \begin{cases} 1 + x & \text{if } -1 < x < 0 \\ 1 - x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that

$$f_{X,Y}(x, y) \neq f_X(x)\, f_Y(y).$$

Hence X and Y are not independent. Let us compute Cov (X,Y). It is easy to check that

$$E(X) = 0 \text{ and } E(Y) = 2/3$$

and

$$E(XY) = \int_0^1 \left[\int_{-y}^{y} x\, dx \right] dy$$

$$= \int_0^1 y \left[\frac{x^2}{2} \right]_{-y}^{y} dy$$

$$= \int_0^1 y \left[\frac{y^2}{2} - \frac{y^2}{2} \right] dy$$

$$= 0.$$

Hence

$$\text{Cov}(X, Y) = 0.$$

Remark : This example shows that **Cov (X,Y) = 0 does not imply that X and Y are independent.**

See if you can solve these exercises.

E15) Suppose the random vector (X,Y) has the joint density function given by

$$f(x, y) = \begin{cases} \frac{1}{\pi}, & x^2 + y^2 \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Then find

i) the marginal density functions of X and Y

ii) the covariance of X and Y

iii) $E(X^2)$.

E16) Show that

$$\text{Cov}(aX + b, cY + d) = ac\ \text{Cov}(X, Y)$$

for any constants a,b,c,d.

Next we shall define moment generating functions for a bivariate random vector.

Suppose (X,Y) has a bivariate density function $f_{X,Y}(x, y)$.

As in the univariate case the moment generating function of (X,Y) is defined as

$$M_{(X,Y)}(t_1, t_2) = E\left[e^{t_1 X + t_2 Y}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x + t_1 y} f_{X,Y}(x, y)\, dx\, dy \qquad \ldots(14)$$

whenever the integral exists. Here t_1 and t_2 are real numbers. Let us write $M(t_1, t_2)$ for $M_{X,Y}(t_1, t_2)$ for simplicity. You can check that

$$M(0,0) = 1,$$

$$M(t_1, 0) = E\left[e^{t_1 X}\right] = M_X(t_1), \text{ and}$$

$$M(0, t_2) = E\left[e^{t_2 Y}\right] = M_Y(t_2).$$

In other words, the moment generating functions of the marginal distributions of X and Y can be recovered from the moment generating function of the joint distribution of (X,Y).

Now let us see how to generate the moments of a bivariate (X,Y) from its moment generating function. For that we assume that we can differentiate the expression $e^{t_1 x + t_1 y}$, r times with respect to t_1 and s times with respect to t_2 where $r \geq 0$ and $s \geq 0$. Then from (14) we have

$$\frac{\partial^{r+s} M(t_1,t_2)}{\partial t_1^{\,r}\, \partial t_2^{\,s}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^r\, y^s\, e^{t_1 x + t_2 y} f(x,y)\, dx\, dy$$

Let us see some particular cases of this expression. For $r = 1, s = 0, t_1 = 0 = t_2$, we have

$$\left.\frac{\partial M(t_1,t_2)}{\partial t_1}\right|_{\substack{t_1=0,\\ t_2=0}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dx\, dy = E(X) = \mu_X$$

Similarly we have

$$\left.\frac{\partial M(t_1,t_2)}{\partial t_2}\right|_{\substack{t_1=0,\\ t_2=0}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f_{X,Y}(x,y)\, dx\, dy = E(y) = \mu_Y$$

When $r = 2, s = 0, t_1 = 0 = t_2$, we have

$$\left.\frac{\partial^2 M(t_1,t_2)}{\partial t_1^{\,2}}\right|_{\substack{t_1=0,\\ t_2=0}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f_{X,Y}(x,y)\, dx\, dy = E(X^2)$$

and when $r = 0, s = 2, t_1 = 0 = t_2$, we have

$$\text{n}\ \frac{\partial^2 M(t_1, t_2)}{\partial t_2^{\,2}}\Bigg|_{\substack{t_1=0,\\ t_2=0}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f_{X;Y}(x,y)\, dx\, dy = E(Y^2)$$

We also have

$$\left.\frac{\partial^2 M(t_1, t_2)}{\partial t_1\, \partial t_2}\right|_{\substack{t_1 = 0,\\ t_2 = 0}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\, dx\, dy = E(XY).$$

Now, do you agree with me when I say that

$$\left.\frac{\partial^{r+s} M(t_1,t_2)}{\partial t_1^r\, \partial t_2^s}\right|_{\substack{t_1 = 0,\\ t_2 = 0}} = E(x^r y^s)?$$

closely look at the particular cases, then you won't have much problem to this fact.

above discussion it follows that the moments $E(X^r Y^s)$ if it exists, can be rom $M(t_1,t_2)$ by differentiating the same partially with respect to t_1 and t_2 d number of times. You know that this is the reason why $M(t_1,t_2)$ is called m generating function.

portant special case is the case of the moment generating function when X and independent random variables. Let us look at this problem.

X and Y are independent random variables with the density functions $f_x(x)$ respectively. Then the joint density function of (X , Y) is $f_X(x)\, f_Y(y)$ and

$$\begin{aligned} M(t_1,t_2) &= E\left[e^{t_1 X + t_1 Y}\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x + t_2 y} f_X(x)\, f_Y(y)\, dx\, dy \\ &= \left\{\int_{-\infty}^{\infty} e^{t_1 x} f_X(x)\, dx\right\}\left\{\int_{-\infty}^{\infty} e^{t_2 y} f_Y(y)\, dy\right\} \\ &= E\left[e^{t_1 x}\right] E\left[e^{t_2 y}\right] \\ &= M_X(t_1)\, M_Y(t_2). \end{aligned}$$

sider an example now.

Example : Suppose (x , y) has the joint density function

$$\begin{aligned} f(x,y) &= e^{-y}, \quad 0 < x < y < \infty \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

Let us compute $M(t_1, t_2)$

$$M(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x + t_2 y} f(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty}\left\{\int_{x}^{\infty} e^{t_1 x + t_2 y} e^{-y} dy\right\} dx$$

$$= \int_{0}^{\infty}\left\{\int_{x}^{\infty} e^{-(1-t_2)y}\, dy\right\} e^{t_1 x}\, dx$$

$$= \int_{0}^{\infty}\left[\frac{e^{-(1-t_2)y}}{-(1-t_2)}\right] e^{t_1 x} \quad \text{for } t_2 < 1$$

$$= \frac{1}{1-t_2}\int_{0}^{\infty} e^{-(1-t_2-t_1)x}\, dx \quad \text{for } t_2 < 1$$

$$= \frac{1}{(1 - t_1 - t_2)(1 - t_2)}$$

for $t_1 < 1$ and $t_1 + t_2 < 1$. In particular

$$M(0, t_2) = \frac{1}{(1 - t_2)^2}, \quad t_2 < 1$$

is the m.g.f. of X.

In this section we have discussed many concepts for a bivariate random vector. We have talked about expectation, variance, covariance, moments and moment generating functions. Among these concepts, you have seen that covariance enables us to measure the strength of association between two random variables. However covariance is not a good measure of the relationship as it depends on the units of measurements for X and Y. In the next section we shall discuss two more concepts which measures the relationship between the random variables.

12.6 CORRELATION AND REGRESSION

In Unit 4 , we have already introduced the concept of correlation coefficient between two variables X and Y. You may recall that the correlation coefficient between X and Y is defined as

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\{\text{Var}(X)\ \text{Var}(Y)\}}}$$

where in the denominator , the +ve square root is taken. The same definition holds good in the case of bivariate continuous random vectors (X, Y) as well.

You have already seen in Unit 4 that if $\rho_{x,y}$ is the correlation coefficient between X and Y then $-1 \leq \rho_{X,Y} \leq 1$. We can give a direct proof of this result.

Before we start proving that we shall introduce you to an important inequality known as Cauchy-Schwartz Inequality. This inequality is a useful tool in probability theory in various contexts.

Suppose (W, Z) is a bivariate random vector and let us consider the function

$$g(W, Z) = (W - kZ)^2$$

where k is a real number. From the general properties of expectation. We have

$$\begin{aligned} 0 \leq E[(W - kZ)^2] &= E[W^2 - 2kWZ + k^2Z^2] \\ &= E[W^2] - 2kE(WZ) + k^2E(Z^2). \end{aligned}$$

The last expression is a quadratic function in k. Since this quadratic function is non-negative for all real k, the discriminant of the quadratic function has to be negative. Therefore

$$4[E(WZ)]^2 - 4E(W^2)E(Z^2) \leq 0$$

or equivalently

$$[E(WZ)]^2 \leq E(W^2)E(Z^2) \qquad \ldots(15$$

This inequality is known as **Cauchy-Schwartz inequality**. If equality occurs in the above inequality, then it follows that the discriminant is zero and hence the quadratic equation

$$k^2E(Z^2) - 2kE(WZ) + E(W^2) = 0$$

has equal roots. In other words, there exists a value k_0 such that

$$k_0^2 E(Z^2) - 2k_0E(WZ) + E(W^2) = 0$$

that is

$$E\left[(W - k_0 Z)^2\right] = 0.$$

Utilizing the Cauchy-Schwartz inequality, we can prove the following theorem.

Theorem 2 : Let (X , Y) be a bivariate random vector with $E(X) = \mu_X$, $E(Y) = \mu_Y$, $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$ and $Cov(X, Y) = \sigma_{X,Y}$. Then $|\rho_{X,Y}| \leq 1$.

Proof : Choose

$$W = X - \mu_X \text{ and } Z = Y - \mu_Y$$

in the earlier discussion. Then, we have $\rho_{X,Y}^2 \leq 1$ . ...(16)

Next we shall prove another theorem.

Theorem 3 : Suppose that $\rho^2 = 1$. Then Y is a linear function of X.

Proof : Let us consider relation (16).

As we have seen in the earlier discussion the equality occurs in the relation (16) iff there exists some value $k_0, (k_0 \neq 0)$ such that

$$P\left[X - \mu_X = k_0(Y - \mu_Y)\right] = 1$$

that is $Y = \frac{1}{k_0}\left[\mu_Y - \mu_X + X\right]$ with probability one. In other words Y is a linear function of X or equivalently X and Y are linearly related with probability 1 .

Does the converse of this theorem hold ? That is, if X and Y are two random variables such that Y is a linear function of X, then is it true that $\rho^2 = 1$?

You can find an answer to this question in the following theorem.

Theorem 4 : Suppose that X and Y are two random variables for which $Y = a X + b$ where a and b are constants. Then $\rho^2 = 1$. In fact if $a > 0$, then $\rho = +1$ and if $a < 0$, $\rho = -1$.

Why don't you try to prove this theorem by yourselves ? (see E 17).

Now let us consider the bivariate random vector discussed in Example 8. What is $\rho_{X,Y}$ for that random vector ? $\rho_{x,y} = 0$, since we have shown that cov (X , Y) = 0. Such random variables X and Y are said to be uncorrelated. More precisely, two random variables X and Y are said to be **uncorrelated** if the correlation coefficient between X and Y is zero.

We have proved in the last section that if X and Y are independent random variables with finite variance, then they are uncorrelated since cov(X, Y) = 0 and hence $\rho_{x,y} = 0$. However as mentioned in the remark below Example 7, it is **not true** that if the random variables are uncorrelated, then they are independent.

Before we go any further, it is time to do some exercises now.

E17) Prove Theorem 4. [Hint : $E(Y) = aE(X) + b$ and $V(Y) = a^2 V(X)$. Compute E (X Y) and then $\rho_{X,Y}^2$].

E18) Suppose that (X, Y) is a bivariate random vector with the joint density function

$$f_{X,Y}(x,y) = \begin{cases} 8xy, & \text{if } 0 < x < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Compute the correlation coefficient.

E19) Suppose U is a random variable uniformly distributed over $[0, 2\pi]$. Define $X = \cos U$ and $Y = \sin U$. Show that X and Y are uncorrelated.

E20) If X has standard normal distribution and $Y = a + bX + cX^2$, $b \neq 0$, or $c \neq 0$, then show that

$$\rho_{X,Y} = \frac{-b}{\sqrt{b^2 + 2c^2}}$$

If you have done these exercises, you would have got a fairly good grasp of a correlation coefficient. Next we shall talk about concept of Regression which gives the relationship between the variable .

In Unit 4, we have already introduced the concept of regression curve. Recall that locus of the conditional expectation of Y given X = x is called the regression of Y on X. Similarly, you may recall from Unit 4 that locus of the conditional expectation of X given Y = y is the regression of X on Y. More specifically, suppose (X, Y) is a bivariate random vector with joint density function $f_{X,Y}(x, y)$. The conditional density of X given Y = y is defined by

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad -\infty < x < \infty$$

for all y such that $f_Y(y) > 0$. Similarly the conditional density of Y given X = x is defined by

$$f_{X|Y}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \quad -\infty < y < \infty$$

The **conditional expectation** of g (Y) given X = x is denoted by $E[g(Y) \mid X = x]$ and it is equal to

$$\int_{-\infty}^{\infty} g(y)\, f_{Y|X}(y \mid x)\, dy.$$

In other words, it is just the expectation of g (Y) with respect to the conditional density of Y given X = x. Similarly we define

$$E\left[h(X) \mid Y = y \right]$$

$$= \int_{-\infty}^{\infty} h(x)\, f_{X|Y}(x \mid y)\, dx.$$

Of course, all these expectations make sense only when the corresponding integrals are finite. If we choose g (y) = y and h (x) = x , then we have

$$E\left[Y \mid X = x \right] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy$$

and

$$E\left[X \mid Y = y \right] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, dx.$$

Let us consider the conditional expectation of Y given X = x.

Note that E(Y | X = x) depends on the value x for X. Hence it is a function depending on X = x. Denote this function by Q(x). The function Q(x) = E (Y | X = x) is called **regression** or **regression function** of Y on X. If Q is linear function (say) a + bx for some constants a and b , then Y is said to have **linear regression** on X.

Similarly if E (X | Y = y) is a linear function of y, then X is said to have linear regression on y and Q(y) = E [X | Y = Y] is called the regression of X on Y. For simplicity , we will denote E (Y | X) for Q(x) and E (X | Y) for Q (y).

Let us consider an example .

Example 11 : Suppose (X, Y) has the joint density function

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

Let us compute the regression of Y on X and that of X on Y and show that both are linear.

$$f_X(x) = \int_x^1 2\,dy = 2(1 - x) \quad \text{for } 0 < x < 1$$

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

and

$$f_Y(y) = \int_0^y 2\,dx = 2y \text{ for } 0 < y < 1$$

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

Hence

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x, y)}{f_Y(x)} = \frac{1}{y} \text{ for } 0 < x < y, \quad 0 < y < 1$$

$$= 0 \quad \text{elsewhere}$$

and

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x, y)}{f_X(x)} = \frac{1}{1 - x} \quad \text{for } x < y < 1, 0 < x < 1$$

$$= 0 \quad \text{elsewhere}$$

Therefore, the regression of Y on X ,

$$E(Y \mid X = x) = \int_{-\infty}^{+\infty} y\, f_{Y|X}(y \mid x)\, dy \quad \text{for } 0 < x < 1$$

$$= \frac{1 + x}{2} \qquad \text{for } 0 < x < 1$$

and the regression of X on Y is

$$E(X \mid Y = y) = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, dx$$

$$= \int_0^y x \frac{1}{y}\, dx \qquad \text{for } 0 < y < 1$$

$$= \frac{y}{2} \qquad \text{for } 0 < y < 1$$

Hence the regression of Y on X is linear as well as the regression of X on Y is linear.

Now, if the regression of Y on X is linear , say $E(X \mid Y) = ax + b$, then we can express the coefficients a and b in terms of certain parameters of the joint distribution of (X, Y). The same is true for $E(X \mid Y)$ also. The following theorem illustrates this point.

Theorem 5 : Let (X , Y) be a random vector and suppose that $E(X) = \mu_X$, $E(Y) = \mu_Y$) , $V(X) = \sigma_X^2$ and $V(Y) = \sigma_Y^2$ Let ρ be the correlation coefficient between X and Y. If the regression of Y on X is linear , we have

$$E(Y \mid X) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X).$$

Similarly if the regression of X on Y is linear , we have

$$E(X \mid Y) = \mu_X + \rho\frac{\sigma_Y}{\sigma_X}(y - \mu_Y).$$

You can try to prove this theorem by yourself (see E24).

We shall now make some remarks on Theorem 1.

Remark : 1) If the regression of X on Y is linear and if $\rho = 0$, then relation (16) shows that $E(X \mid Y)$ does not depend on y. Similarly if the regression of Y on X is linear and if $\rho = 0$, then $E(Y \mid X)$ does not depend on x.

2) Relation (16) also shows that if the regression of Y on X is linear , then $E(Y \mid X)$ represents a straight line whose slope is $\rho\frac{\sigma_X}{\sigma_Y}$. Therefore the sign of ρ determines the slope of the regression line.

Now, here are some exercises which you should try to solve.

E21) Consider the joint probability density function

$$f(x, y) = \begin{cases} 1 & \text{if } -y < x < y,\ 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

of a bivariate random vector (X, Y). Are both the regressions linear ?

E22) Prove that

$$E(E(Y \mid X)) = E(Y)$$

whenever the expectations exist.

E23) Suppose (X , Y) has the joint probability density function

$$f(x, y) = y^2 e^{-y(x+1)}, \quad x \geq 0,\ y \geq 0$$
$$= 0 \quad \text{otherwise}$$

Find $E(Y \mid X)$.

E24) If (X, Y) is a bivariate random vector such that $E(Y \mid X = x)$ is a linear function of x, then show that

$$E(Y \mid X = x) = \mu_Y + \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X^2 = \text{Var}(X)$, $\sigma_Y^2 = \text{Var}(Y)$ and $\rho_{X,Y}$ is the correlation coefficient between X and Y.

E25) Show that if X and Y are independent, then

$$E(X \mid Y = y) = E(X)$$

for all y and $E(Y \mid X = x) = E(Y)$ for all x.

Now we bring this unit to a close. But before that let's recall the important concepts that we studied in it.

12.7 SUMMARY

In this unit we have covered the following points :

1) We have introduced you to the notion of a bivariate distribution and the associated concepts of joint density function.

2) We have acquainted you with the concepts of conditional distribution : The conditional density function of X given Y = y is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, -\infty < x < \infty$$

for any y such that $f_Y(y) > 0$

Similarly the conditional density function of Y given X = x is defined as

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, -\infty < y < \infty$$

for any x such that $f_X(x) > 0$.

3) We have defined and discussed the consequence of independence of two events.

4) We have generalized the notions of expectation, moments and moment generating function for a bivariate distribution.
If g (X, Y) is a function of the random vector (X, Y), then

$$E[g(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\, f_{X,Y}(x,y)\, dx\, dy$$

where $f_{X,Y}(x,y)$ is the joint density function.

The product moment $= E[(x-a)^r (y-b)^s]$

$$\text{Var}(X) = E[(x-a)^2]$$

$$\text{Var}(Y) = E[(y-b)^2]$$

$$\text{Cov}(X,Y) = E[(x-a)(y-b)].$$

The moment generating function of (X, Y) is

$$M_{X,Y}(t_1,t_2) = E\left[e^{t_1x_1 + t_2x_2}\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{t_1x_1 + t_2x_2} f_{X,Y}(x,y)\, dx\, dy$$

5) We have investigated the concepts of Correlation coefficient and Regression between the variables.

Correlation coefficient $\rho_{X,Y} = \dfrac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$

Regression of Y on X $= E[Y|X]$.

Regression of X on Y $= E[X|Y]$.

12.8 SOLUTIONS AND ANSWERS

E1) The required probability is $P[X \geq 2Y] = \int_0^{\infty}\int_0^{x/2} xy\, e^{-(x+y)}\, dy\, dx$

$$= \int_0^{\infty} x\, e^{-x}\left[\int_0^{x/2} y\, e^{-y}\, dy\right] dx$$

$= 7/27.$

E2) $P[X + Y > 500] = 1 - P[X + Y \leq 500]$

$$= 1 - \int_0^{\infty}\int_0^{500-x} \lambda^2 e^{-\lambda(x+y)}\, dy\, dx$$

$$= 1 - \{1 - e^{-500\lambda} - 500\lambda\, e^{-500\lambda}\}$$

$$= e^{-500\lambda} + 500\lambda\, e^{-500\lambda}$$

$$= e^{-500\lambda}(1 + 500\lambda).$$

E3) By definition, $f_{X,Y}(x, y) = \dfrac{\partial^2 F_{X,Y}(x, y)}{\partial x\, \partial y}$

$$= \frac{\partial^2 [(1 - e^{-\lambda y})(1 - e^{-\lambda x})]}{\partial x\, \partial y}, \text{ if } x > 0, y > 0$$

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

$$\therefore \quad f_{X,Y}(x, y) = \lambda^2 e^{-\lambda(x+y)} \text{ if } x > 0, y > 0$$

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

E4) $f_X(x) = \dfrac{2}{(x+1)^3}, \quad x > 0$

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

$$f_Y = ye^{-y}, \ y > 0$$

$$= 0, \quad y \le 0.$$

E5) The expression (q) in (6) can be rewritten as

$$q = \frac{\left[x - \left\{\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y - \mu_y)\right\}\right]^2}{\sigma_x^2(1 - \rho^2)} + \frac{[\mu - \mu_y]^2}{\sigma_y^2}$$

Then we get

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2}\left(\frac{y - \mu_Y}{\sigma_Y}\right)^2}$$

Therefore the marginal density of Y is normal with mean μ_Y and variance σ_Y^2

E6) By definition

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

From the property

$$\int_{-\infty}^{\infty}\int_{\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = \int_0^1\int_0^1 c(x + y^2)\, dx\, dy = 1$$

We get that c = 1.

$$\therefore \quad f_{X,Y}(x, y) = (x + y^2), \text{ if } 0 < x, y < 1$$

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

Also we have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx$$

$$= \int_0^1 (x + y^2)\, dx$$

$$= \frac{x^2}{2} + y^2 x \Big]_0^1$$

$$= \frac{1}{2} + y^2), \quad 0 < x < 1$$

$$\therefore \quad f_{X|Y}(x \mid y) = \frac{x + y^2}{\frac{1}{2} + y^2}, \quad 0 < x < 1:$$

Now,

$$P\left[X < \frac{1}{2} \mid Y = \frac{1}{2}\right] = \int_0^{1/2} f_{X|Y=1/2}(x \mid y = 1/2)\, dx$$

$$= \frac{1}{3}.$$

E7) Here $f_X(x) = 2e^{-2x}, \quad x > 0$

$= 0, \quad x \leq 0$

and

$$f_{Y|X}(y \mid x) = e^{x-y}, \; y > x$$

$= 0$ otherwise

Hence

$$P[Y < 1 \mid X < 1] = \frac{P[X < 1, Y < 1]}{P[X < 1]}.$$

Now,

$$P[X < 1, Y < 1] = P[X < Y, Y < 1]$$

$$= 1 - 2e^{-1} + e^{-2}$$

and

$$P[X < 1] = 1 - e^2$$

Hence

$$P[Y < 1 \mid X < 1] = \frac{1 - 2e^{-1} + e^{-2}}{1 - e^{-2}}.$$

Further

$$P[Y < 1 \mid X = 1] = \int_0^1 f_{Y|X=1}(y \mid x = 1)\, dy =$$

Hence

The joint density function is $f_{X,Y}(x, y) = \frac{2y + 4x}{3}, \quad 0 < x, y < 1$

$= 0,$ otherwise.

E8) $f_Y(y) = \frac{2y + 2}{3}$ if $0 < y < 1$

$= 0$ otherwise.

E9) Here

$$f_X(x) = \frac{2}{5}\left(2x + \frac{3}{2}\right), \quad 0 < x < 1$$

$= 0$, otherwise

and

$$f_Y(y) = \frac{2}{5}(1 + 3y), \quad 0 < y < 1$$

$= 0$ otherwise

Hence, for $0 < y < 1$,

$$f_{X|Y}(x \mid y) = \frac{2x + 3y}{1 + 3y}, \quad \text{if } 0 < x < 1$$

$= 0,$ otherwise

and, for $0 < x < 1$,

$$f_{Y|X}(y \mid x) = \frac{2x + 3y}{2x + 3/2} \quad \text{if } 0 < y < 1$$

$= 0$ otherwise

Then

$$P\left[X > \frac{4}{5}\right] = \int_{4/5}^{1} f_X(x)\,dx,$$

$$P\left(X > \frac{4}{5} \mid Y = \frac{3}{10}\right) = \int_{4/5}^{1} f_{X|Y}(x \mid y = 3/10)\,dx,$$

and

$$P\left(Y > \frac{4}{5} \mid X = \frac{3}{10}\right) = \int_{4/5}^{1} f_{Y|X}(y \mid x = 3/10)\,dy.$$

These probabilities can be computed using the functions given above.

E10) Suppose that the events are independent. Then we have

$$F_{X,Y}(x, y) = F_X(x)\,F_Y(y), -\infty < x, y < \infty$$

Therefore

$$\frac{\partial^2 F_{X,Y}(x, y)}{\partial x\ \partial y} = \frac{\partial^2}{\partial x\ \partial y}\left[F_X(x)\,F_Y(y)\right]$$

$$\frac{\partial}{\partial x}\left[F_{(X)}(x)\right]\frac{\partial}{\partial y}\left[F_Y(y)\right]$$

$$= f_X(x)\,f_Y(y)$$

E11) Here

$$f_X(x) = 2x \quad \text{if } 0 < x < 1$$
$$= 0 \quad \text{otherwise}$$

and

$$f_Y(y) = 6y(1 - y) \quad \text{if } 0 < y < 1$$
$$= 0 \quad \text{otherwise}$$

Further more

$$f(x, y) = f_X(x)\,f_Y(y)$$

for all x and y. Hence x and y are independent random variables.

E12) Here

$$f_X(x) = 2x \quad \text{if } 0 < x < 1$$
$$= 0 \quad \text{otherwise}$$

and

$$f_Y(y) = 2(1 - y) \quad \text{if } 0 < y < 1$$

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

Since $f(x, y) = f_X(x)\,f_Y(y)$ for all x and y, X and Y are independent. Therefore

$$P\left[0 < X < \frac{1}{3}, 0 < Y < \frac{1}{3}\right] = P\left[0 < X < \frac{1}{3}\right] P\left[0 < Y < \frac{1}{3}\right]$$

$$= \int_0^{1/3} 2x\,dx \int_0^{1/3} 2(1 - y)\,dy$$

$$= \frac{1}{9} \times \frac{10}{18} = \frac{5}{81}.$$

E13) X and Y are independent.

E14) $\text{Cov}(x, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

$[XY - \mu_X Y - \mu_Y x + u_x \mu_Y]$

$$= E[XY] - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y$$
$$= E(XY) - \mu_X \mu_Y.$$
$$= E(XY) - E(X)\,E(Y).$$

E15) i) $f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\, dy$

$$= 2\int_0^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy$$

$$= \frac{2}{\pi}\sqrt{1-x^2},\ 0 < x < 1.$$

Similarly

$$f_Y(y) = \frac{2}{\pi}\sqrt{1-y^2},\ 0 < y < 1$$

ii) $\text{Cov}(XY) = E(XY) - E(X)\,E(Y)$

Now,

$$E(X) = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx$$

$$= 0$$

and

$$E(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\, y\, f_{X,Y}(x, y)\, dx\, dy$$

This integral can be easily evaluated using polar coordinates. Recall that you have studied evaluation of double integrals using polar coordinates in MTE-07, Block 4. Then we have

$$E(XY) = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} r^3 \sin\theta \cos\theta\, d\theta$$

$$= \frac{1}{\pi}\int_0^1 r^3 \left[-\frac{\cos 2\theta}{4}\right]_0^{2\pi} d\theta$$

$$= 0.$$

$\therefore$ $\text{Cov}(XY) = 0.$

E16) Observe that $E(aX + b) = aE(X) + b$ and $E(cY + d) = cE(Y) + d$

Hence $\text{Cov}(aX + b, cX + d) = E\left[\{(aX + b) - E(aX + b)\}\{cY + d - E(cY + d)\}\right]$

$$= E\left[a\{X - E(X)\}\, c\{Y - E(Y)\}\right]$$

$$= ac\,\text{Cov}(X, Y).$$

E17) Since $Y = aX + b$, we have $E(Y) = aE(X) + b$ and

$\text{Var}(Y) = a^2\,\text{Var}(X)$. Also

$$E(XY) = E[X(aX + b)]$$

$$= E(aX^2 + b)$$

$$= aE(X^2) + b$$

Therefore

$$\rho^2 = \frac{[E(XY) - E(X)\,E(Y)]^2}{\text{Var}(X)\text{Var}(Y)}$$

$$= \frac{[aE(X^2) + bE(X) - E(X)(aE(X) + b)]^2}{a^2\,\text{Var}(X)\,\text{Var}(X)}$$

$$= \frac{[aE(X^2) + bE(X) - aE(X)^2 - bE(X)]^2}{a^2\text{Var}(X)^2}$$

$$= a^2 \text{Var}(X)^2$$

$$= \frac{a^2 \text{ Var } (X)^2}{a^2 \text{ Var } (X)^2} = 1.$$

E18) From Example 8, we know that $E(X) = \frac{8}{15}$, $E(Y) = \frac{4}{5}$ and $\text{cov}(XY) = \frac{4}{9} - \frac{32}{75}$.

Now to calculate Var (X) and Var (Y)

$$\text{Var } (X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = \int_0^1 \int_0^y x^2 \, 8xy \, dx \, dy$$

$$= \int_0^1 8y \, dy \left[\int_0^y x^3 \, dx \right.$$

$$= 2 \int_0^1 y^5 \, dy$$

$$= \frac{1}{3}$$

$$\therefore \quad \text{Var } (X) = \frac{1}{3} - \frac{64}{225} = \frac{11}{225}.$$

Similarly we get that $\text{Var } (Y) = \frac{2}{75}$.

$$\rho_{(X,Y)} = \frac{\text{cov } (X, Y)}{\sqrt{\text{Var } (X) \text{ Var } (Y)}}$$

$$= \frac{4/9 - 32/75}{\sqrt{\frac{11}{225} \times \frac{2}{75}}}$$

$$= \frac{4}{\sqrt{66}}.$$

E19) The density function f_U of U is

$$f_U(u) = c, \text{ if } u \in [0, 2\pi]$$

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

where c is a constant.

Then we get

$$E(X) = \int_{-\infty}^{+\infty} \cos u \, f_u(u) \, du = 0,$$

$$E(Y) = 0$$

and

$$E(XY) = E[\sin u \cos u] = \int_0^{2\pi} \sin u \cos u \, f_U(u) \, du = 0$$

$$\therefore \quad \text{cov } (X, Y) = 0$$

and hence

$$\rho_{X,Y} = 0$$

E20) Note that $E(X) = 0$ and $\text{Var } (X) = 1$

Now

$$E(Y) = E[a + bX + cX^2]$$

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2$$
$$= E[a^2 + b^2X^2 + c^2X^4 + 2abX + 2abcX^3 + 2acX^2] - (a + c)^2$$
$$= b^2 + 2c^2$$

and

$$\text{Cov}(X,Y) = E(XY) = b$$

$$\therefore \quad \rho_{X,Y} = \frac{b}{\sqrt{b^2 + 2c^2}}$$

E21) From the calculation made in Example 9 , we have

$$E(X \mid Y = y) = \int_{-y}^{y} x \frac{1}{2y}\, dx \quad \text{for } 0 < y < 1.$$
$$= 0 \qquad \text{for } 0 < y < 1.$$

On the other hand

$$E[Y \mid X = x] = \int_{-x}^{1} y.\frac{1}{1+x}\, dy \text{ if } -1 < x < 0$$
$$= \int_{x}^{1} y \cdot \frac{1}{1-x}\, dy \text{ if } 1 < x < 0$$

Hence

$$E(Y \mid X = x) = \begin{cases} \frac{1-x}{2} & \text{if } -1 < x < 0 \\ \frac{1+x}{2} & \text{if } 0 < x < 1 \end{cases}$$

Therefore X has linear regression on Y but Y does not have linear regression on X.

$$E(Y \mid X) = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy$$

E22) $$E[E(Y \mid X)] = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy \right] f_X(x)\, dx$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, f_X(x)\, dy\, dx$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f_{X,Y}(x, y)\, dy\, dx$$
$$= E(Y).$$

E23) $$E(Y \mid X) = \frac{3}{x+1}$$

E24) We discuss the case when (X , Y) has a joint probability density function. We leave it to you to check the result when (X,Y) has a bivariate discrete distribution.

Now suppose (X,Y) is a continuous distribution. From E22 it follows that $E[E(Y \mid X)] = E(Y) = \mu_Y$

But, by hypothesis, $E(Y \mid X)$ is a linear function of X i.e. we can write

$$E(Y \mid X) = a + bX.$$

where a,b are real numbers such that $b \neq 0$.

Hence

$$\mu_Y = E[E(Y \mid X)] = E(a + bX) = a + b\,\mu_X$$

Therefore

$$a = \mu_Y - b\,\mu_X$$

and

$$E(Y \mid X) = \mu_Y - b\,\mu_X + bX$$

i.e. $$E(Y \mid X) - \mu_Y = b(X - \mu_X).$$

Multiplying both sides by $(X - \mu_X)$ and taking expectation on both sides, we get that

$$E[[E(Y \mid X) - \mu_Y][X - \mu_X]] = bE[X - \mu_X]^2$$

$$\{E[E(Y \mid X)] - \mu_Y\} E[(X - \mu_X)] = bE[X - \mu_X]^2$$

$$\{E[Y] - \mu_Y\} E[(X - \mu_X)] = bE[X - \mu_X]^2$$

$$E[Y - \mu_Y] E[X - \mu_X] = bE[x - \mu_X]^2.$$

Verify that the relation leads to

$$E[(Y - \mu_Y)(X - \mu_X)] = bE[X - \mu_X]^2.$$

In other words

$$\text{Cov}(X,Y) = b\,\sigma_X^2$$

or

$$b = \frac{\text{Cov}(X,Y)}{\sigma_X^2} = \frac{\rho_{X,Y}\,\sigma_Y}{\sigma_X}$$

Hence

$$E(Y \mid X) = a + bX$$

$$= \mu_Y - b\,\mu_X + bX$$

$$= \mu_X + \frac{\rho_{X,Y}\,\sigma_Y}{\sigma_X}(X - \mu_X).$$

E25) Suppose X and Y are independent with density functions $f_X(x)$ and $f_Y(y)$ respectively. Then

$$E(X \mid Y = y) = \int_{-\infty}^{\infty} x\, f_{X \mid Y}(x \mid y)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{x\, f_{X \mid Y}(x, y)}{f_Y(y)}\, dx$$

$$= \int_{-\infty}^{\infty} f_X(x)\, dx$$

$$= E(X).$$

Verify the result for discrete distributions.

UNIT 13 FUNCTIONS OF RANDOM VARIABLES

Structure

13.1 INTRODUCTION

In the last unit, we have introduced bivariate distributions and multivariate distributions. Most of the times we would like to know the probabilistic behaviour of a function g(X,Y) of the random vector (X,Y). The function g could be either the sum X+Y or the max (X,Y) or some other function depending on the phenomenon under study. In Section 13.2, we give two approaches for obtaining the distribution function of a function of two random variables. Important distributions such as Chi-square distribution, t-distribution and F-distribution are studied in sections 12.3 to 12.5. These distributions can be considered as distributions of functions of independent standard normal random variables. Properties of these distributions are investigated in detail.

Objectives

After reading this unit, you should be able to :

- derive distribution functions of functions of two or more random variables ;
- derive properties of the Chi-square, t and F = distributions;
- explain the connection between Chi-square, t and F and the normal distributions.

13.2 FUNCTION OF TWO RANDOM VARIABLES

In this section we shall talk about functions of two random variables and discuss methods for obtaining their distribution functions. Some of the important functions which we shall consider are X+Y, XY, $\frac{X}{Y}$, max (X,Y), |X-Y|.

Let us start with a random vector (X,Y). By definition X and Y are random variables defined on the sample space S of some experiment and each of which assigns a real number to every $s \in S$. Let g(x,y) be a real-valued function defined on $\mathbf{R} \times \mathbf{R}$. Then the composite function Z = g(X,Y) defined by

$$Z(s) = g[X(s), Y(s)], s \in S$$

assignes to every outcome $s \in S$ a real number. Z is called a **function of the random** vector (X,Y).

For example, if $g(x,y) = x + y$, then we get $Z = X + Y$ and if $g(x,y) = xy$, then we get $Z = XY$ and so on.

Now let us see how do we find the distribution function of Z. As in the univariate case, we shall restrict ourselves to the continuous case. Here we shall discuss two methods for obtaining distribution functions – Direct Method and Transformation Method. We shall first discuss Direct Method.

13.2.1 Direct Method

Let (X,Y) be a random vector with the joint density function $f_{X,Y}(x,y)$. Let $g(x,y)$ be a real-valued function defined on $\mathbf{R} \times \mathbf{R}$. For $z \in Z$, define

$$D_z = \{(x, y) : g(x,y) \leq z\}$$

Then the distribution function of Z is defined as

$$P[Z \leq z] = \int_{D_z} \int f_{X,Y}(x,y)\, dx\, dy \qquad \ldots(1)$$

Theoretically it is not difficult to write down the distribution function using (1). But in actual practise it is sometimes difficult to evaluate the double integral.

We shall now illustrate the computation of distribution functions in the following examples.

Example 1: Suppose (X,Y) has the uniform distribution on $[0,1] \times [0,1]$ the unit square. Then the joint density of (X,Y) is

$$f_{X,Y}(x,y) = \begin{cases} 1 & \text{if } 0 < x, y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Let us find the distribution function of $Z = g(X,Y) = XY$.

From the definition of a distribution function of Z, we have

$$F_Z(z) = P[XY \leq z]$$
$$= \int_{D_z} \int f_{X,Y}(x,y)\, dxdy \quad \text{if } 0 < z < 1$$

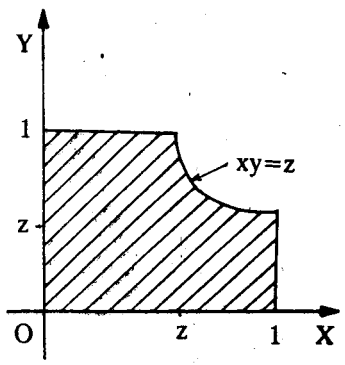


Fig. 1

where $D_z = \{(x,y) : \ldots \leq z\}$

$$= \iint_{D_z^1} dx\, dy \quad \text{if } 0 < z < 1.$$

where $D_z^1 = \{(x,y) : xy \leq z, 0 < x < 1, 0 < y < 1\}$

In order to evaluate the last integral, let us look at the set of all points (x,y) such that $xy \leq z$ when $0 < x < 1$ and $0 < y < 1$ (See Fig. 1).

If $0 < x < z$, then for any $0 < y < 1$, the product $xy \leq z$ and if $x > z$, then $xy \leq z$ only when $0 < y < z/x$. This is the region shaded in Fig. 1. Hence for $0 < z < 1$

$$F_Z(z) = P[Z \leq z]$$

$$= \int_0^z \left\{ \int_0^1 dy \right\} dx + \int_z^1 \left\{ \int_0^{z/x} dy \right\} dx$$

$$= \int_0^z dx + \int_z^1 \frac{z}{x}\, dx$$

$$= z + z\,[\ln x]_z^1 = z - z \ln z.$$

Therefore

$$F_Z(z) = \begin{cases} 0 & \text{if } z = 0 \\ z - z \ln z & \text{if } 0 < z < 1 \\ 1 & \text{if } z \geq 1 \end{cases}$$

is the distribution function of Z. The density function $f_Z(z)$ of Z is obtained by differentiating $F_Z(z)$ with respect to z. Then you can check that

$$f_Z(z) = 0 \quad \text{if } z \leq 0 \text{ or } z \geq 1$$

$$= -\ln z \quad \text{if } 0 < z < 1$$

Example 2: Suppose X and Y are independent exponential random variables with the density function

$$f(x) = \lambda e^{-\lambda x}, x > 0 \quad \text{and } f(y) = \lambda e^{-\lambda y}, y > 0$$

$$= 0, \quad x \leq 0. \qquad = 0, y \leq 0$$

Define $Z = X + Y$ and let us find the distribution function of Z.

From the definition of Z,

$$F_Z(z) = P[Z \leq z] = 0 \quad \text{if } z < 0$$

and, for $z > 0$

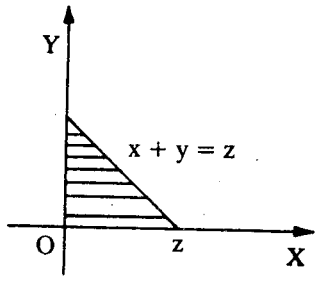


Fig. 2

$$F_Z(z) = P[Z \le z]$$

$$= \int\int_{\{(x,y): x+y \le z\}} f_{X,Y}(x,y)\,dx\,dy$$

where $f_{X,Y}(x, y)$ is the joint density of (X,Y). Since X and Y are independent random variables, the joint density of (X,Y) is given by

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

where $f_X(x)$and $f_Y(y)$ are the marginal density functions

i.e. $$f_{X,Y}(x, y) = \lambda e^{-\lambda x} \times \lambda e^{-\lambda y} \quad , x > 0, y > 0$$

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

Now for $z > 0$, the set $\{(x, y) : x + y \le z, x > 0, y > 0\}$ is the region shaded in Fig. 2. Hence, for $z > 0$,

$$F_Z(z) = \int_0^z \left[\int_0^{z-x} \lambda e^{-\lambda y}\,dy \right] \lambda e^{-\lambda x}\,dx$$

$$= \int_0^z \left[-e^{-\lambda x} \right]_0^{z-x} \lambda e^{-\lambda x}\,dx$$

$$= \int_0^z \left[1 - e^{-\lambda(z-x)} \right] \lambda e^{-\lambda x}\,dx$$

$$= \int_0^z \left[\lambda e^{-\lambda x} - \lambda e^{-\lambda x} \right] dx$$

$$= \left[-e^{-\lambda x} \right]_0^z - z\lambda e^{-\lambda z}$$

$$= 1 - e^{-\lambda z} - \lambda z e^{-\lambda z}.$$

Now we leave it to you to check that the density function of Z is

$$f_Z(z) = \lambda^2 z\, e^{-\lambda z} \qquad \text{for } z > 0$$

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

In this density function familiar to you? Recall that this function is the gamma density function you have studied in Unit 11. Hence Example 2 says that the sum of two independent exponential random variables has gamma distribution.

Let us consider another example.

Example 3 : Suppose X and Y are independent random varibles with the same density function f (x) and the distribution function F(x). Define Z = max(X,Y). Let us determine the distribution function of Z.

By definition, the distribution function F_Z is given by

$$F_Z(z) = P[Z \le z]$$

$$= P[\max(X, Y) \le z]$$

$$= P[X \le z, Y \le z]$$

$$= P[X \le z]\, P[Y \le z] = [F(z)]^2$$

by the independence of X and Y and the fact that

$$P[X \le z] = P[Y \le z] = F(z).$$

Since F is differentiable almost everywhere and the density corresponding to F is f it follows that Z has a probability density function f_Z and

$$f_Z(z) = 2F(z)\, f(z), \quad -\infty < z < \infty .$$

To get more practise why don't you try some exercises now.

E1) Suppose X and Y are independent random variables, each having uniform distribution on (0, 1). Determine the density function of $Z = X + Y$.

E2) Suppose (X,Y) has the joint probability density function

$$f(x,y) = x + y, \text{ if } 0 < x, y < 1$$
$$= 0, \text{ otherwise}$$

Find the density function of $Z = XY$.

E3) Suppose X and Y are independent r. vs with density function f(x) and distribution function F(x). Find the density function of $Z = \min(X,Y)$.

The examples and exercises discussed above deal with the method of obtaining the distribution function of $Z = g(X,Y)$ directly. This method is applicable even when (X,Y) does not have a density function.

Next we shall discuss another method for obtaining the distribution and density functions.

13.2.2 Transformation Approach

Suppose (X_1, X_2) is a bivariate random vector with the density function $f_{X_1, X_2}(x_1, x_2)$ and we would like to determine the distribution function of the density function of $Z_1 = g_1(X_1, X_2)$. To determine this, let us suppose that we can find another function $Z_2 = g_2(X_1, X_2)$ such that the transformation from (X_1, X_2) to (Z_1, Z_2) is one-to-one. In other words to every point (x_1, x_2) in $\mathbf{R}^2$, there corresponds a point (z_1, z_2) in $\mathbf{R}^2$ given by the above transformation and conversely to every point (z_1, z_2) there corresponds a unique point (x_1, x_2) such that

$$z_1 = g_1(x_1, x_2)$$
$$z_2 = g_2(x_1, x_2)$$

For example suppose that $Z_1 = X_1 + X_2$. Then we can choose $Z_2 = X_1 - X_2$. You can easily see that the transformation $(x_1, x_2) \to (z_1, z_2)$ from $\mathbf{R}^2$ to $\mathbf{R}^2$ is one-one and in this case we have

$$X_1 = \frac{Z_1 - Z_2}{2} \text{ and } X_2 = \frac{Z_1 + Z_2}{2}$$

So, in general, one can assume that we can express (X_1, X_2) in terms of (Z_1, Z_2) uniquely.

That means that there is exist real valued functions h_1 and h_2 such that

$$X_1 = h_1(Z_1, Z_2)$$
$$X_2 = h_2(Z_1, Z_2)$$

Let us further assume that h_1 and h_2 have continuous partial derivatives with respect to Z_1, Z_2. Consider the Jacobian of the tranformation $(Z_1, Z_2) \to (X_1, X_2)$

$$\begin{vmatrix} \frac{\partial h_1}{\partial z_1} & \frac{\partial h_1}{\partial z_2} \\ \frac{\partial h_2}{\partial z_1} & \frac{\partial h_2}{\partial z_2} \end{vmatrix} = \frac{\partial h_1}{\partial z_1}\frac{\partial h_2}{\partial z_2} - \frac{\partial h_1}{\partial z_2}\frac{\partial h_2}{\partial z_1}$$

Recall that you have seen 'Jacobians' in Unit 9, Block 3 of MTE - 07 we denote this Jacobian by $J = \frac{\partial(x_1, x_2)}{\partial(z_1, z_2)}$. Assume that J is not zero for all (z_1, z_2). Then, it can be shown, by using the change of variable formula for double integrals [see MTE-07, Unit 11, Block 4] we can show that the random vector (Z_1, Z_2) has a density and the density function $\phi(z_1, z_2)$ of (Z_1, Z_2) is

$$\phi(z_1, z_2) = f\left[h_1(z_1, z_2), h_2(z_1, z_2)\right] |J| \text{ if } (z_1, z_2) \in B \qquad \ldots(2)$$
$$= 0 \qquad \text{otherwise}$$

where $B = \{(z_1, z_2) : z_1 = g_1(x_1, x_2), z_2 = g(x_1, x_2) \text{ for some } (x_1, x_2)\}$.

From the joint density of (Z_1, Z_2) obtained above, the marginal density of Z_1 can be derived and it is given by

$$\phi_2(z_1) = \int_{-\infty}^{\infty} \phi(z_1, z_2)\, dz_2$$

Let us now compute the density function of $Z_1 = X_1 + X_2$ where X_1 and X_2 are independent and identically distributed standard normal variables. We have seen that in this case $Z_2 = X_1 + X_2$ and we can write

$X_1 = \frac{Z_1 + Z_2}{2}$ and $X_2 = \frac{Z_1 - Z_2}{2}$. Let us now calculate the Jacobian of the transformation. It is given by

$$\begin{vmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Now since X_1 and X_2 are independent, we have

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_2^2}, -\infty < x_1, x_2 < \infty.$$

Hence by (1) the joint probability density function of (Z_1, Z_2) is

$$\phi(z_1, z_2) = f\left[\frac{z_1 + z_2}{2}, \frac{z_1 - z_2}{2}\right] |J|, -\infty < z_1, z_2 < \infty$$

$$= \frac{1}{4\pi} \exp\left\{-\frac{1}{2}\left[\frac{z_1 + z_2}{2}\right]^2 - \frac{1}{2}\left[\frac{z_1 - z_2}{2}\right]^2\right\}$$

$$= \frac{1}{4\pi} \exp\left\{-\left[\frac{z_1^2}{4} + \frac{z_2^2}{4}\right]\right\}, -\infty < z_1, z_2 < \infty$$

Then the marginal density of Z_1 is given by

$$\phi_{Z_1}(z_1) = \int_{-\infty}^{\infty} \phi(z_1, z_2)\, \partial z_2 \frac{1}{\sqrt{4\pi}} e^{\frac{z_1^2}{4}}, -\infty < z_1 < \infty$$

Note that we can calculate the marginal density of Z_2 also. It is given by

$$\phi_{Z_2}(z_2) = \int_{-\infty}^{+\infty} \phi(z_1, z_2)\, \partial z_1 = \frac{1}{\sqrt{4\pi}} e^{-\frac{z_2^2}{4}}, -\infty < z_2 < \infty.$$

In other words Z_1 has $N(0, 2)$ and Z_2 has $N(0, 2)$ as their distribution functions. In fact Z_1 and Z_2 are independent random variables since

$$\phi(z_1, z_2) = \phi_{Z_1}(z_1)\, \phi_{Z_2}(z_2)$$

for all z_1 and z_2.

We shall illustrate this method with one more example.

Example 4 : Suppose X_1 and X_2 are independent random variables with common density function

$$f(x) = \frac{1}{2} e^{-x/2} \quad \text{for } 0 < x < \infty$$
$$= 0 \quad \text{otherwise.}$$

Let us find the distribution function of $Z_1 = \frac{1}{2}(X_1 - X_2)$.

Here it is convenient to choose $Z_2 = X_2$. Note that the transformation $(X_1, X_2) \to (Z_1, Z_2)$ gives a one-to-one mapping from the set $A = \{(x_1, x_2) : 0 < x_1 < \infty, 0 < x_2 < \infty\}$ onto the set $B = \{(z_1, z_2) : z_2 > 0, -\infty < x_2 < \infty \text{ and } z_2 > -2z_1\}$. The inverse transformation is

$$X_1 = 2Z_1 + Z_2$$

and

$$X_2 = Z_2.$$

Since $x_1 > 0$, it follows that $2z_1 + z_2 > 0$, that is, $z_2 > -2z_1$.

Since $x_2 > 0$, it follows that $z_2 > 0$. Obviously, $-\infty < z_1 < \infty$.

Further more you can check that the Jacobian of the transformation is equal to 2.

Now the joint density function is

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{4} e^{\frac{-(x+y)}{2}}$$

Therefore from (2) the joint density function of (Z_1, Z_2) is

$$\phi[z_1, z_2] = f_{X_1, X_2}[2z_1 + z_2, z_2] |J| \quad \text{provided } (z_1, z_2) \in B$$
$$= 0 \quad , \quad \text{otherwise.}$$

Therefore

$$\phi[z_1, z_2] = \frac{1}{2} e^{-z_1 - z_2} \quad \text{if } (z_1, z_2) \in B$$
$$= 0 \quad \text{otherwise} \quad \cdots *$$

and the marginal density of Z_1 is

$$\phi_{Z_1}(z_1) = \int_{-2z_1}^{\infty} \frac{1}{2} e^{-z_1 - z_2} dz_2 \quad \text{if } -\infty < z_1 < 0$$
$$= \int_{0}^{\infty} \frac{1}{2} e^{-z_1 - z_2} dz_2 \quad \text{if } 0 \le z_1 < \infty.$$

This shows that

$$\phi_{Z_1}(z_1) = \frac{1}{2} e^{-|z_1|} \quad , -\infty < z_1 < \infty$$

The distribution with the density function given by * is known as double exponential distribution.

An important application of the transformation approach is to determine distribution of the sum of two independent random variables not necessarily identically distributed. Let us now look at this problem.

Suppose X_1 and X_2 are independent random variables with the density functions $f_1(x_1)$ and $f_2(x_2)$ respectively and we want to determine the distribution function of $X_1 + X_2$. Let $Z_1 = X_1 + X_2$. We apply transformation method here. Set $Z_2 = X_2$. Then the transformation $(X_1, X_2) \to (Z_1, Z_2)$ is invertible and

$$X_1 = Z_1 - Z_2$$
$$X_2 = Z_2.$$

The Jacobian of the transformation is equal to unity. Since the joint density of (X_1,X_2) is $f_1(x_1)\, f_2(x_2)$, it follows from (2) that the joint density of (Z_1, Z_2) is given by

$$\phi(z_1, z_2) = f_1(z_1 - z_2)\, f_2(z_2), \quad -\infty < z_1, z_2 < \infty.$$

Hence the density of Z_1 is given by

$$\phi_{Z_1}(z_1) = \int_{-\infty}^{\infty} \phi(z_1, z_2)\, dz_2$$

$$\int_{-\infty}^{\infty} f_1(z_1 - z_2)\, f_2(z_2)\, dz_2 \quad , -\infty < z_1, z_2, < \infty \qquad \ldots(3)$$

This formula giving the density function of Z_1 is known as the **convolution formula.**

This is called the convolution formula because the density function is the convolution product of the density functions of X_1 and X_2.

The convolution product of two real-valued functions f_1 and f_2, denoted by $f_1 * f_2$, is defined by

$$f_1 * f_2(x) = \int_{-\infty}^{\infty} f_1(x - t)\, f_2(t)\, dt.$$

Let us now calculate the distribution function of Z_1. We denote the distribution function of Z_1 by ϕ_{Z_1}. Then we have

$$\phi_{Z_1}(z) = \int_{-\infty}^{z} \phi_{Z_1}(z_1)\, dz_1$$

$$= \int_{-\infty}^{z} \left[\int_{-\infty}^{\infty} f_1(z_1 - z_2)\, f_2(z_2)\, dz_2 \right] dz_1$$

Since f_1 and f_2 are continuos, we can interchange the order of integration (see MTE-07, Block 4 Unit 11)

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z} f_1(z_1 - z_2)\, dz_1 \right] f_2(z_2)\, dz_2$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z - z_2} f_1(u)\, du \right] f_2(z_2)\, dz_2$$

(by the transformation $u = z_1 - z_2$)

$$= \int_{-\infty}^{\infty} F_1(z - z_2)\, f_2(z_2)\, dz_2$$

where F_1 is the distribution function of X_1.

Therefore the distribution function of Z_1 is the convolution product of the distribution function of X_1 and the density function of X_2.

The above relation gives an explicit formula for the distribution function of Z_1.

Let us see an example.

Example 5 : Suppose X_1 and X_2 are independent random variables with the gamma distributions having parameters (α_1, λ) and (α_2, λ) respectively. Let us find the density function of the sum $Z = X_1 + X_2$ using the convolution formula.

The density of X_1 is

$$f_{X_i}(x_i) = \frac{\lambda^{\alpha_i} x_i^{\alpha_i - 1} e^{-\lambda x_i}}{\Gamma(\alpha_i)}, \qquad x_i > 0$$

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

for i = 1, 2. We use Formula (3) to compute the density function of 2. For $z > 0$, we have

$$\phi_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(z-u)\, f_{X_2}(u)\, du$$

$$= \int_0^{\infty} f_{X_1}(z-u)\, f_{X_2}(u)\, du$$

$$= \int_0^z \frac{\lambda^{\alpha_1} e^{-\lambda(z-u)}}{\Gamma(\alpha_1)} (z-u)^{\alpha_1 - 1} \frac{\lambda^{\alpha_2} e^{-\lambda u}}{\Gamma(\alpha_2)} u^{\alpha_2 - 1}\, du .$$

$$= \frac{\lambda^{\alpha_1+\alpha_2} e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^z \left\{ (z-u)^{\alpha_1 - 1} u^{\alpha_2 - 1}\, du \right\}$$

$$= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\lambda z} z^{\alpha_1+\alpha_2-1} \left\{ \int_0^1 (1-v)^{\alpha_1-1} v^{\alpha_2-1}\, dv \right\}$$

(by the transformation $v = \frac{u}{z}$)

$$= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\lambda z} z^{\alpha_1+\alpha_2-1} B(\alpha_2, \alpha_1)$$

$$\therefore \phi_Z(z) = \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)} e^{-\lambda z} z^{\alpha_1+\alpha_2-1}, \quad 0 < z < \infty .$$

$$= 0 \quad , \quad z < 0.$$

The last equality follows from the properties of beta function and gamma function.

This example shows that the convolution of gamma distributions with parameters (α_1, λ) and (α_2, λ) is a gamma distribution with parameter $(\alpha_1 + \alpha_2, \lambda)$.

Next we shall consider another example in which we illustrate another method called Moment Generating Function approach. This method is useful for finding the distribution functions of sums or linear combinations of independent random variables.

Example 6 : Suppose X_1 and X_2 are independent random variables with distributions $N[\mu_1, \sigma_1^2]$ and $N[\mu_2, \sigma_2^2]$ respectively. Define $Z = X_1 + X_2$. Then the m.g.f. of Z is

$$M_Z(t) = E\left[e^{t(X_1+X_2)}\right]$$

$$= E\left[e^{tX_1} e^{tX_2}\right]$$

$$= E\left[e^{tX_1}\right] E\left[e^{tX_2}\right].$$

The last relation follows from the fact that e^{tX_1} and e^{tX_2} are independent random variables when X_1 and X_2 are independent. But we have proved earlier that

$$E\left[e^{tX_i}\right] = \exp\left\{\mu_i t + \frac{1}{2} t^2 \sigma_i^2\right\}, \text{ i.= 1, 2}$$

Hence

$$M_Z(t) = \exp\left\{t\left[\mu_1 + \mu_2\right] + \frac{1}{2} t^2 \left[\sigma_1^2 + \sigma_2^2\right]\right\}, \quad -\infty < t < \infty .$$

But this function is the m.g.f. of $N[\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2]$. From the uniqueness property (Theorem 1 of Unit 10), it follows that Z has $N[\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2]$.

Have you understood the method discussed in this section? To verify that why don't you try some exercises now.

E4) Suppose X_1 and X_2 are independent random variables with gamma densities $f_i(x_i)$ given by

$$f_i(x_i) = \frac{1}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} e^{-x_i}, \quad 0 < x_i < \infty$$
$$= 0 \quad \text{otherwise}$$

for i = 1, 2. Let $Z_1 = X_1 + X_2$ and $Z_2 = \frac{X_1}{X_1 + X_2}$. Show that Z_1 and Z_2 are independent random variables. Find the distribution functions of Z_2 and Z_1.

E5) **(Box - Muller transformation)** Let X_1 and X_2 be independent random variables uniformly distributed on [0, 1]. Define

$$Z_1 = (-2 \log X_1)^{1/2} \cos(2\pi X_2),$$
$$Z_2 = (-2 \log X_1)^{1/2} \sin(2\pi X_2)$$

Show that Z_1 and Z_2 are independent standard normal random variables.

E6) Suppose (X_1, X_2) have the joint density function

$$f(x_1, x_2) = 4x_1 x_2 \quad \text{if } 0 < x_1, x_2 < 1$$
$$= 0 \quad \text{otherwise.}$$

Define $Z_1 = \frac{X_1}{X_2}$ and $Z_2 = X_1 X_2$. Determine the joint density function of (Z_1, Z_2).

E7) Suppose $X_1, \ldots, X_n$ are n independent random variables with the same distribution $N(\mu, \sigma^2)$. Define

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$\overline{X}$ is called the sample mean. Extending the m.g.f. approach for more than two random variables, show that $\overline{X}$ has the distribution $N\left[\mu, \frac{\sigma^2}{n}\right]$.

In the next section we shall talk about functions of more than two random variables.

13.3 FUNCTIONS OF MORE THAN TWO RANDOM VARIABLES

Suppose we have n random variables $X_1, \ldots, X_n$ not necessarily independent and we are interested in finding the distribution function of a function $Z_1 = g_1(X_1, \ldots, X_n)$ or the joint distribution function of $Z_i = g_i(X_1, \ldots, X_n)$, $1 \le i \le r$, where r is any positive integer $1 \le r \le n$. The methods described in the previous section can be extended to this general case. We will not go into detailed description or extension of the methods. We will illustrate by a few examples.

Example 7 : Suppose $X_1, X_2, \ldots X_n$ is a random sample of size n, from a certain population. We shall discuss this concept of random sampling in greater detail in Block 4 Unit 15. In the present context it will suffice to record that the above statement is a convenient alternative way of expressing the fact that $X_1, X_2, \ldots X_n$ are independent and identically distributed n random variables with a common distribution function F(x) which coincide with the population distribution function

(see Unit 15, Block 4). Define $Z_1 = \min(X_1,.....,X_n)$ and $Z_n = \max(X_1,.....,X_n)$. Let us find the joint distribution of (Z_1, Z_n).

We first note that $Z_1 \le Z_n$. Let us compute the distribution function G_{Z_1, Z_n} of (Z_1, Z_n). Let (z_1, z_n) be a fixed pair where $-\infty < z_1 \le z_n < \infty$. We first consider the case $z_1 = z_n$. Then

$$\begin{aligned} G_{Z_1, Z_n}(z_1, z_n) &= P[Z_1 \le z_1, Z_n \le z_n] \\ &= P[Z_n \le z_n], \text{ since the event } [Z_n \le z_n] \\ &\qquad \text{implies the event } [Z_1 \le z_1], \\ &= P[X_i \le Z_n \ 1 \le i \le n] \quad z_1 \text{ and } z_n \text{ being equal.} \\ &= \prod_{i=1}^{n} P[X_i \le z_n], \text{ since } X_i\text{'s are independent} \\ &= F(z_n)^n. \end{aligned}$$

Now, suppose that $z_1 < z_n$. Then we have

$$\begin{aligned} G_{Z_1, Z_n}(z_1, z_n) &= P[Z_1 \le z_1, Z_n \le z_n] \\ &= P[Z_n \le z_n] - P[Z_n < z_n, Z_1 > z_1] \\ &= P[Z_n \le z_n] - P[z_1 < Z_1 \le Z_n \le z_n] \\ &= P[Z_n \le z_n] - P(z_1 < X_i \le z_n \text{ for } 1 \le i \le n) \\ &= P(Z_n \le z_n) - \prod_{i=1}^{n} P[z_1 < X_i \le z_n], \text{since } X_i\text{'s are independent} \\ &= P[X_i \le z_n \text{ for } 1 \le i \le n] - \prod_{i=1}^{n} P[z_1 < X_i \le z_n] \\ &= \prod_{i=1}^{n} P[X_i \le z_n] - \prod_{i=1}^{n} P[z_1 < X_i \le z_n] \\ &= [F(z_n)]^n - [F(z_n) - F(z_1)]^n. \end{aligned}$$

Therefore if $-\infty < z_1 \le z_n < \infty$, we get the distribution function as

$$G_{Z_1, Z_n}(Z_1, Z_n) = F(Z_n)^n - [F(Z_n) - F(Z_1)]^n \qquad \ldots(4)$$

The joint probability density function of (Z_1, Z_n) is obtained by the relation

$$g_{Z_1, Z_n}(z_1, z_n) = \frac{\partial^2 G_{Z_1, Z_n}(z_1, z_n)}{\partial z_1 \, \partial z_n}.$$

Then from (4), we have

$$\begin{aligned} g_{Z_1, Z_2}(z_1, z_n) &= n(n-1)[F(z_n) - F(z_1)]^{n-2} f(z_1) f(z_n) \text{ if } -\infty < z_1 < z_n < \infty \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

The quantity $Z_n - Z_1$ is called the **range**. Infact, Range is the difference between the largest and the smallest observations. We shall now find the distribution of the range $W_1 = Z_n - Z_1$ for the observations given in Example 7.

Example 8 : Let $X_1, X_2, ..., X_n$ and Z_1, Z_n are as given Example 7. Let us find the distribution of $W_1 = Z_n - Z_1$.

Here we make use of the transformation method.

Set $W_2 = Z_1$

Now you can check that the transformation $(Z_1, Z_n) \to (W_1, W_2)$ is one-to-one and the inverse transformation is given by $Z_1 = W_2$, $Z_n = W_1 + W_2$. The Jacobian of this transformation is equal to -1. Hence the joint density of (W_1, W_2) is given by

$$G(w_1, w_2) = g_{Z_1, Z_n}(w_2, w_2+w_1),\ 0 < w_1 < \infty,\ -\infty < w_2 < \infty.$$

where g_{Z_1, Z_n} is the joint density of (Z_1, Z_n) which we have calculated in Example 7. Then we have

$$G(w_1, w_2) = n(n-1)\,[(F(w_2+w_1) - F(w_2)]^{n-2}\, f(w_2) f(w_2+w_1) \text{ if } 0 < w_1 < \infty \text{ and } -\infty < w_2 < \infty$$

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

and the marginal density function of W_1 is

$$\phi_{w_1}(w_1) = \int_{-\infty}^{\infty} \phi(w_1, w_2)\, dw_2$$

$$= n(n-1) \int_{-\infty}^{\infty} [F(w_2 + w_1) - F(w_2)]^{n-2}\, f(w_2)\, f(w_2+w_1)\, dw_2 \text{ If } 0 < w < \infty$$

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

Let us consider a special case of the above problem when $X_1, \ldots, X_n$ are independent and identically distributed with uniform distribution on [0, 1]. Then

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

and

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

In this case

$$\phi_{W_1}(w_1) = n(n-1) \int_0^{1-w_1} w_1^{n-2}\, dw_2 \quad , \text{ if } 0 < w_1 < 1$$

$$= n(n-1) w_1^{n-2} (1-w_1) \quad , \text{ if } 0 < w_1 < 1$$

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

Now for a short exercise

E8) Suppose $X_1, X_2, \ldots, X_n$ are independent random variables and X_i has $N(\mu_i, \sigma_i^2)$ as its distribution for i=1, 2, ..., n. Find the distribution of $Z = \sum_{i=1}^{n} C_i X_i$ where C_i are constants not all zero.

In the next three sections we shall discuss three standard distributions each of which appear as the distribution of a certain function of standard normal variable. We shall make use of the different approaches discussed in this unit to obtain their distribution functions. All these distributions play an important role in statistical inference which will be discussed in Block 4.

13.4 CHI-SQUARE DISTRIBUTION

In this section we shall introduce you to a standard distribution known as chi-square distribution.

Suppose X has the standard normal distribution. Let us compute the distribution function of $Z = X^2$. Then, for $z \geq 0$,

$$F_Z(z) = P[Z \le z]$$
$$= P[X^2 \le z]$$
$$= P[-\sqrt{z} \le X \le \sqrt{z}]$$
$$= P[X \le \sqrt{z}] - P[X < -\sqrt{z}]$$
$$= \phi[\sqrt{z}] - \phi[-\sqrt{z}]$$

where ϕ is the standard normal distribution function.

It is obvious that $F_Z(z) = 0$ for $z < 0$.

Differentiating F_z, we have the probability density function for Z, namely,

$$f_Z(z) = \frac{1}{2} z^{-1/2} \phi(\sqrt{z}) + \frac{1}{2} z^{-1/2} \phi(-\sqrt{z}) \quad \text{for } z \ge 0.$$
$$= 0 \quad \text{for } z < 0$$

where

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, -\infty < u < \infty.$$

Hence

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z/2} z^{-1/2}, 0 \le z < \infty$$
$$= 0 \quad \text{otherwise.}$$

Another equivalent form of the density function is

$$f_Z(z) = \frac{1}{2^{1/2}\Gamma\left(\frac{1}{2}\right)} z^{(1/2)-1} e^{-z/2} \quad \text{for } z \ge 0.$$
$$= 0 \quad \text{for } z < 0$$

χ is a Greek letter pronounced as 'kai'.

A distribution with the above density function is called a **Chi-square distribution with 1 degree of freedom** and it is denoted by χ^2. We have now proved that if X is a standard normal random variable, then X^2 has a χ^2 distribution with 1 degree of freedom. Let us now compute the m.g.f. of $Z = X^2$.

By the definition of m.g.f. it follows that

$$M_Z(t) = E[e^{tZ}]$$
$$= \int_0^\infty e^{tz} \frac{1}{2^{1/2}\Gamma(1/2)} z^{(1/2)-1} e^{-z/2} dz$$
$$= \frac{1}{2^{1/2}\Gamma\left(\frac{1}{2}\right)} \int_0^\infty z^{(1/2)-1} e^{-z/2(1-2t)} dz$$
$$= \frac{1}{2^{1/2}\Gamma\left(\frac{1}{2}\right)} \int_0^\infty u^{-1/2} e^{-u/2} \frac{1}{(1-2t)^{1/2}} du$$

(by the transformation $z(1-2t) = u$)

provided that $t < \frac{1}{2}$.

Note that the integrals are finite only when $t < \frac{1}{2}$.

But

$$\int_0^\infty u^{-1/2} e^{-u/2} du = 2^{1/2}\Gamma\left(\frac{1}{2}\right)$$

from the property of the gamma function (see Sec. 11.3 of Unit 11) or equivalently from observing that

$$\int_{-\infty}^{\infty} f_Z(z)\,dz = 1.$$

Hence

$$M_Z(t) = \frac{1}{(1-2t)^{1/2}} \text{ for } t < 1/2. \qquad \ldots(5)$$

Let us now suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution with mean zero and variance 1. Consider the function

$$S = X_1^2 + \ldots\ldots + X_n^2.$$

S is the sum of the squares of n independent and identically distributed (i.i.d) standard normal random variables. Let us find the m.g.f. of S^2. Note that

$$\begin{aligned} M_S(t) &= E[\exp(tS)] \\ &= E[\exp(t\{X_1^2 + \ldots\ldots + X_n^2\})] \\ &= \prod_{i=1}^{n} E[\exp(tX_i^2)] \end{aligned}$$

by the independence of X_i, $1 \le i \le n$. But each X_i is N(0, 1). Therefore X_i^2 follow χ^2–distribution with m.g.f given by (5). Hence

$$\begin{aligned} M_S(t) &= \prod_{i=1}^{n} \left(\frac{1}{1-2t}\right)^{1/2} \quad \text{for } t < 1/2 \\ &= \frac{1}{(1-2t)^{n/2}} \quad \text{for } t < 1/2. \end{aligned}$$

We leave it to you to check that the function

$$M(t) = \frac{1}{(1-2t)^{n/2}} \quad \text{for } t < 1/2$$

is the m.g.f. corresponding to the density function

$$\begin{aligned} f(z) &= \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{(n/2)-1} e^{-z/2} \quad \text{for } z > 0 \qquad \ldots(6) \\ &= 0 \quad \text{for } z \le 0. \end{aligned}$$

E9) Determine the moment generating function of the χ^2 distribution with n degrees of freedom directly from the definition of the m.g.f.

The distribution corresponding to the above probability density function given by (6) is called the **Chi–Square distribution with n degrees of freedom.** An application of the uniqueness theorem for m.g.f.'s (Theorem 1 of Unit 10) proves that

$$S = X_1^2 + \ldots\ldots + X_n^2$$

has chi–square distribution with n degrees of freedom. This distribution is usually denoted by χ_n^2. Let us now calculate the mean and variance of the Chi–square distribution with n degrees of freedom. Since

$$S = X_1^2 + \ldots\ldots + X_n^2$$

where $X_1, X_2, \ldots, X_n$ are i.i.d. N(0, 1) random variables, it follows that

$$E(S) = E(X_1^2) + \ldots\ldots + E(X_n^2)$$

and

$$\text{Var}(S) = \text{Var}(X_1^2) + \ldots\ldots + \text{Var}(X_n^2).$$

But $E(X_i^2) = 1$ and $Var(X_i^2) = E(X_i^4) - E(X_i^2) = 2$ for the standard normal random variable X_i for $1 \le i \le n$ (see Unit 11, Sec. 11.2). Hence

$$E(S) = n \text{ and } Var(S) = 2n.$$

Now suppose a random variable Y_i is $N(\mu_i, \sigma_i^2)$ for $1 \le i \le n$ and Y_i, $1 \le i \le n$ are independent. Then, you can check that

$$X_i = \frac{Y_i - \mu_i}{\sigma_i}, \quad 1 \le i \le n$$

are independent $N(0, 1)$ and hence

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{\sigma_i^2} = \sum_{i=1}^{n} X_i^2$$

has a Chi–square distribution with n degrees of freedom. In particular if $\mu_i = \mu$ for all i and $\sigma_i^2 = \sigma^2$ for all i, then $Y_1, Y_2, \ldots, Y_n$ is a random sample from $N(\mu, \sigma^2)$ and

$$S = \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2}$$

has chi–square distribution with n degrees of freedom.

From the density of chi–square distribution for n degrees of freedom given in (6), you might have noticed that it is a special case of the gamma density with $\alpha = n/2$ and $\lambda = 1/2$. The graphs given in Fig. 3 shows the shape of the function chi–square density function for n = 1, 2, 3 and 4.

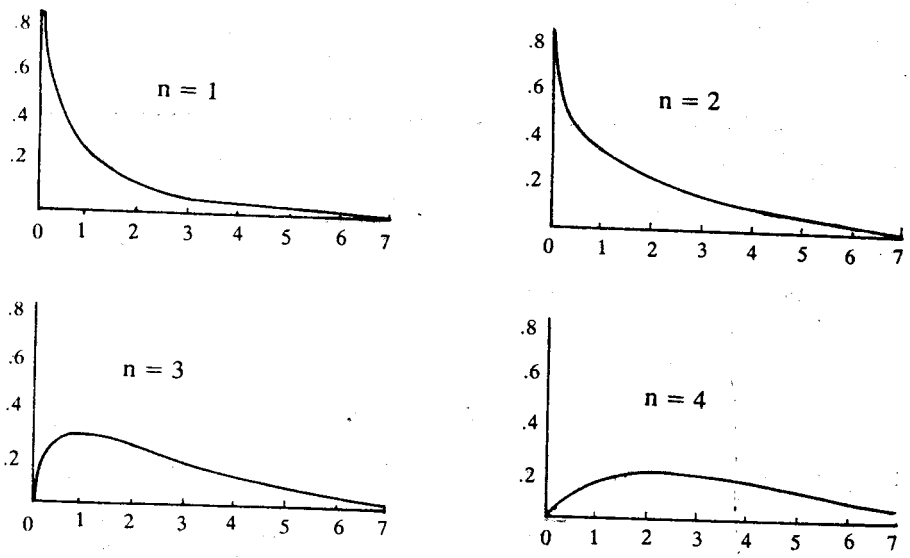


Fig. 3

In general, exact computation of probabilities under χ_n^2–distribution for different values of n is not possible. Tables giving probabilities for the χ_n^2–distribution for different values of n are available. One such table is given at the end of this unit (see appendix). Let us now see how can we compute probabilities under χ_n^2–distribution using the table.

Suppose we want to compute $P[3.25 \le Z \le 20.5]$ when Z has χ_{10}^2 distribution.

It is easy to see from the tables that $P[Z \le 20.5] = 0.975$ and $P[Z \le 3.25] = 0.025$. Hence

$$P[3.25 \le Z \le 20.5] = 0.975 - 0.025 = 0.95$$

an important property of chi–square distribution is the additivity property. We shall illustrate the property in the following theorem.

Theorem 1 : If Z_1 has χ^2_n and Z_2 has χ^2_m–distribution respectively and Z_1 and Z_2 are independent, then Z_1+Z_2 has χ^2_{m+n} distribution.

Proof : Since Z_1 has χ^2_n, Z_1 can be written as the sum of squares of n i.i.d. random variables Y_i, $1 \le i \le n$ each of which is N(0, 1). Similarly Z_2 can be written as the sum of squares of m i.i.d. random variables W_j, $1 \le j \le m$ each of which is N(0, 1).

Hence $Z_1+Z_2 = \sum_{i=1}^{n} Y_i^2 + \sum_{j=1}^{m} W_j^2$ which is the sum of squares of m+n i.i.d. random variables each of which is N(0, 1). Therefore $Z_1 + Z_2$ has χ^2_{m+n} distribution.

Another important result dealing with the chi–square distribution and the normal distributions is given by following theorem. We omit the proof of the result as it is beyond the scope of this course.

Theorem 2 : Suppose $Y_1, \ldots, Y_n$ is a random sample from $N(\mu, \sigma^2)$ with $n > 1$. Let

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

($\overline{Y}$ is the sample mean and S^2 is the sample variance.) Then $\overline{Y}$ and S^2 are independent random variables. $\overline{Y}$ has $N\left(\mu, \frac{\sigma^2}{n}\right)$ as its distribution and $\frac{(n-1)S^2}{\sigma^2}$ has χ^2_{n-1} as its distribution.

This result has a large number of applications in statistical inference as you will see in Block 4.

It is time to do some exercises now.

E10) Show that if Z_1 is χ^2_n and Z_2 is χ^2_m and Z_1 and Z_2 are independent, then $Z_1 + Z_2$ is χ^2_{n+m} by using the m.g.f. approach.

E11) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population with exponential density function

$$f(x) = \lambda e^{-\lambda x}, \; x > 0$$
$$= 0 \quad \text{otherwise.}$$

Show that $Z = 2\lambda n \overline{X}$ has χ^2_{2n} – distribution.

E12) Suppose $X_1, \ldots, X_n$ are independent random variables and X_i has an absolutely continuous distribution function F_i. Define $Y = -2\sum_{i=1}^{n} \log F_i(X_i)$.

Show that Y has χ^2_{2n} – distribution.

(**Hint** :– use the fact that $F_i(X_i)$ has uniform distribution on [0, 1].)

In the next section we shall take up another distribution which is a function of a chi–square distribution and a normal distribution.

13.5 t–DISTRIBUTION

Consider two independent random variables Y and S^2 where Y is N(0, 1) and S^2 is χ^2_n. Define

$$U = \frac{Y}{\left(\frac{S^2}{n}\right)^{1/2}} = \frac{\sqrt{n}\,Y}{\sqrt{S^2}}.$$

The distribution of the random variable U is called **t–distribution with n degrees of**

freedom. The t–distribution is also known as **Student's t–distribution**. The exact derivation of the density function for t is beyond the scope of this book. It can be shown that the density of U is

$$f(u) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{(n\pi)^{1/2}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{u^2}{n}\right)^{-(n+1)/2}, \quad -\infty < u < \infty.$$

It is clear from the form of the density function f(u), that the density function is symmetric about zero. It is bell-shaped and for large n, it is close to the density function for a normal distribution with mean zero.

Now we state some properties of t– distribution. The exact derivation of these properties are beyond the scope of this course.

This distribution does not exist for n = 1. In fact, for n = 1, this distribution is the Cauchy distribution. For n > 1, the mean does exist. Further more, for n > 1, $E(|U|^k) < \infty$ for k < n and $E(|U|^k) = \infty$ for k > n. In other words, the t–distribution with n degrees of freedom with n > 1 has moments up to order n –1 but no moments of higher order exist. In particular, it follows that m.g.f. does not exist for a t–distribution. The typical graphs for various degrees of freedom of the distribution is given in Fig. 4. For computation of probabilities under t–distribution, tables are provided at the end of this Unit (see Appendix). Application of the

This distribution was first obtained by an grish statistician WS gorset, who published his research paper under the name "student"., hence the distribution is also known as student-t- distribution

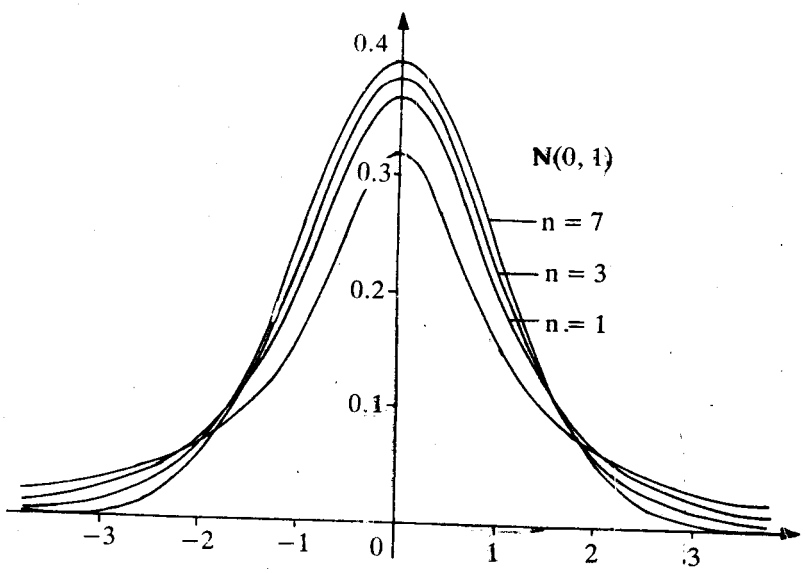


Fig. 4

t–distribution in statistical inference will be treated in Block 4.

You can try some exercises now.

E13) If a random variable U has the t– distribution with n degrees of freedom with n > 2, show that

$$\text{Var}(U) = \frac{n}{n-2}.$$

E14) Suppose a random variable U has the t– distribution with 5 degrees of freedom. Determine $P[1.476 \le U \le 3.365]$.

Apart from χ^2–distribution and t–distribution there is yet another distribution which plays major 'role' in statistical inference. We shall take up that in the next section.

13.6 F-DISTRIBUTION

F–Distribution plays a major role in statistical inference especially in the area of testing of hypothesis. You will see application of this distribution in Block 4. Here we discuss only some major properties of this distribution. Let us first define the F–distribution with m and n degrees of freedom.

Let S_1^2 and S_2^2 be two **independent** random variables such that S_1^2 has $\chi^2_{n_1}$–distribution and S_2^2 and $\chi^2_{n_2}$ – distribution. Define

$$W = \frac{S_1^2/n_1}{S_2^2/n_2^2} = \frac{n_2 S_1^2}{n_1 S_2^2}$$

The distribution of W is called the **F–distribution with n_1 and n_2 degrees of freedom.** This distribution in this particular form is due to Snedecor. It is said that he named this distribution in honour of R.A. Fischer. It can be shown that the density function of W is

$$f(w) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right) n_1^{n_1/2}\, n_2^{n_2/2}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \frac{w^{(n_1/2)-1}}{(n_1 w+n_2)^{\frac{n_1+n_2}{2}}}, \; w > 0.$$
$$= 0, \; w \geq 0.$$

To avoid confusion, sometimes n_1 is referred to as numerator degrees of freedom and n_2 is referred to as denominator degrees of freedom.

You should note that the order (n_1, n_2) is important in defining the F–distribution unless $n_1 = n_2$. The F–distribution with n_1 and n_2 degrees of freedom is entirely different from the F–distribution with n_2 and n_1 degrees of freedom whenever $n_1 \neq n_2$. In fact if W has the F–distribution with n_1 and n_2 degrees of freedom, then 1/W has the F–distribution with n_2 and n_1 degrees of freedom why don't you try this for yourselves (see E16). Let us see some examples.

Example 9 : Suppose $X_1, \ldots, X_m$ is a random sample of size m from $N(\mu, \sigma_1^2)$ and $Y_1, \ldots, Y_n$ is another independent random sample of size n from $N(\mu_2, \sigma_2^2)$ where μ_1, μ_2, σ_1^2 and σ_2^2 are all known. Then $\sum_{i=1}^{m} \frac{(X_i - \mu_1)^2}{\sigma_1^2}$ is χ^2_m

and

$$\sum_{j=1}^{n} \frac{(Y_j - \mu_2)^2}{\sigma_2^2} \text{ is } \chi^2_n.$$

Hence

$$W = \frac{\frac{1}{m}\sum_{i=1}^{m} \frac{(X_i-\mu_1)^2}{\sigma_1^2}}{\frac{1}{n}\sum_{j=1}^{n} \frac{(Y_j-\mu_2)^2}{\sigma_2^2}}$$

has the F– distribution with m and n degrees of freedom.

Example 10. Suppose $X_1, \ldots, X_m$ is a random sample from $N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_n$ is another independent random sample from $N(\mu_2, \sigma^2)$. Observe that both the population have the same variance σ^2. From the remarks made in Example 1, it follows that

$$W = \frac{\frac{1}{m}\sum_{i=1}^{m}\frac{(X_i-\mu_1)^2}{\sigma^2}}{\frac{1}{n}\sum_{j=1}^{n}\frac{(Y_j-\mu_2)^2}{\sigma^2}}$$

has the F– distribution with m and n degrees of freedom. However W does not depend on σ^2 from the above expression. Note that $W = \frac{n\sum_{i=1}^{m}(X_i-\mu_1)^2}{m\sum_{j=1}^{n}(Y_j-\mu_2)^2}$

In other words the distribution of W is F– distribution with m and n degrees of freedom irrespective of the fact whether σ^2 is known or unknown.

Example 11 : Suppose $X_1, ..., X_m$ is a random sample from $N(\mu_1, \sigma_1^2)$ and $Y_1, ..., Y_n$ is another independent random sample from $N\left(\mu_2, \sigma_2^2\right)$ where μ_1, μ_2 and σ_1^2, σ_2^2 are all unknown. Let us consider $S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i-\overline{X})^2$ and

$S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j-\overline{Y})^2$. S_X^2, S_Y^2 are the sample variances of the X-sample and Y-sample respectively. From our discussions in sections 13.2 and 13.4,

$$\frac{(m-1)S_X^2}{\sigma_1^2} \text{ has } \chi^2_{m-1}$$

and

$$\frac{(n-1)S_Y^2}{\sigma_2^2} \text{ has } \chi^2_{n-1}$$

distributions. Hence

$$W = \left\{\left(\frac{1}{m-1}\frac{(m-1)S_X^2}{\sigma_1^2}\right)\left(\frac{1}{n-1}\frac{(n-1)S_Y^2}{\sigma_2^2}\right)\right\}$$

$$= \frac{\sigma_2^2 S_X^2}{\sigma_1^2 S_Y^2}$$

has the F– distribution with (m–1) and (n–1) degrees of freedom. The expression for W involves σ_1^2 and σ_2^2. If $\sigma_1^2 = \sigma_2^2$, then W reduces to

$$W = \frac{S_X^2}{S_Y^2}$$

and W has the F–distribution with (m–1) and (n–1) degrees of freedom.

In order to compute the probabilities under F–distribution for different pairs of degrees of freedom, tables are available. On such table is given at the end of this unit (see Appendix).

See, if you can solve these exercises.

E15) If a random variable has the t–distribution with n degrees of freedom, show that $Z = U^2$ has the F–distributed with 1 and n degrees of freedom.

E16) If W has the F–distribution with m and n degrees of freedom, show that 1/W has the F–distribution with n and m degrees of freedom.

E17) If W has the F–distribution with 7 and 10 degrees of freedom, find a and b such that

$P[W \le a] = 0.975$ and $P[W \le b] = 0.95$.

13.7 SUMMARY

In this unit we have (1) given different approaches for finding the distribution of functions of two or more random variables, and (2) introduced and studied properties of the Chi–square distribution, t–distribution and F–distribution as the distributions of functions of normal random variables.

13.8 SOLUTIONS AND ANSWERS

E1) By definition, the distribution function

$$F_Z(z) = P[Z \le z]$$

$$= P[X + Y \le z]$$

$$= \iint_{D_z} f_{X,Y}(x, y)\, dx\, dy$$

where $D_z = \{x, y\} : x+y \le z\}$ and $f_{x,y}$ is the joint density of (X, Y). Since X and Y are independent r.v's with uniform distribution on [0, 1], the joint density will be given by

$$b_{X,Y}(x, y) = C, \text{ if } 0 \le x, y, \le 1$$

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

where C is a constant.

Therefore we get

$$F_Z(z) = \iint_{D_z} dx\, dy$$

$$= \begin{cases} \frac{z^2}{2} & , \text{ if } 0 < z < 1 \\ 1 - \frac{(2-z)^2}{2} & , \ 1 < z < 2 \\ 0 & , \text{ otherwise} \end{cases}$$

Therefore the p.d.f is

$$f_Z(z) = \begin{cases} z & ; \text{ if } 0 < z < 1 \\ 2-z & , \text{ if } 1 < z < 2 \\ 0 & , \text{ otherwise} \end{cases}$$

E2) $F_Z(z) = \begin{cases} 2 - 2z, & \text{if } 0 < z < 1 \\ 0 & , \text{ otherwise.} \end{cases}$

E3) $F_Z(z) = P[\min(X_1, X_2) \le z]$

$= 1 - P[\min(X_1, X_2) > z]$

$= 1 - P[X_1 > z, X_2 > z]$

$= 1 - P[X_1 > z]\, P[X_2 > z]$ (By the independence of X_1 and X_2)

$= 1 - [\,1 - F(z)]^2$

since X_1 and X_2 are random variables with the same distribution function F(x). Hence, the density of Z is

$$f_Z(z) = 2[1 - F(z)]\, f(z).$$

E4) Note that

$X_1 = Z_1 Z_2$, $X_2 = Z_1 (1-Z_2)$

and the Jacobian of the transformation is

$$J = \begin{vmatrix} z_2 & z_1 \\ 1-z_2 & -z_1 \end{vmatrix} = -z_1 \neq 0$$

in the space $0 < z_1, z_2 < \infty$. Check that the joint density of (Z_1, Z_2) is

$$f_{Z_1, Z_2}(z_1,z_2) = \frac{z_2^{\alpha_1-1}(1-z_2)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z_1^{\alpha_1+\alpha_2-1} e^{-z_1} \text{ for } 0 < z_1 < \infty, 0 < z_2 < 1$$

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

Check that Z_1 and Z_2 are independent random variables. Also verify that

$$f_{Z_2}(z_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z_2^{\alpha_1-1} (1-z_2)^{\alpha_2-1}, \; 0 < z_2 < 1$$

$$= 0 \qquad \text{otherwise.}$$

which is the beta density with parameters α_1 and α_2.

and

$$f_{Z_1}(z_1) = \frac{1}{\Gamma(\alpha_1+\alpha_2)} z_1^{\alpha_1+\alpha_2-1} e^{-z_1}, \; 0 < z_1 < \infty$$

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

which is the gamma density with parameters $\alpha_1 + \alpha_2$ and 1.

E5) Note that

$$X_1 = \exp\left(-\frac{Z_1 + Z_2}{2}\right), X_2 = \frac{1}{2\pi} \tan^{-1} \frac{Z_2}{Z_1}$$

and the Jacobian of the transformation is

$$J = -\frac{1}{2\pi} \exp\left\{-\frac{z_1^2 + z_2^2}{2}\right\}$$

which is not zero. The joint density function of (Z_1, Z_2) is

$$f_{Z_1, Z_2}(z_1, z_2) = \frac{1}{2\pi} e^{-\frac{z_1^2+z_2^2}{2}}, \; -\infty < z_1, z_2 < \infty.$$

Hence Z_1 and Z_2 are independent standard normal random variables.

E6) The joint density function of (Z_1, Z_2) is

$$f_{Z_1, Z_2}(z_1, z_2) = 2\frac{z_2}{z_1} \quad \text{for } z_2 > 0,\, z_1 > 0,\, (z_1 z_2)^{1/2} < 1,\, (z_2/z_1)^{1/2} < 1$$

$$= 0 \qquad \text{otherwise.}$$

E7)
$$E[e^{t\bar{X}}] = E\left[\exp\left\{\frac{t}{n}\sum_{i=1}^{n} X_i\right\}\right]$$

$$= \prod_{i=1}^{n} E\left[\exp\left\{\frac{1}{n} X_i\right\}\right]$$

$$= \exp\left[\frac{t}{n}\mu + \left\{\frac{t}{n}\right\}^2 \frac{\sigma^2}{2}\right]^n$$

$$= \exp\left[t\mu + \frac{t^2\sigma^2}{n}\right]$$

which is the m.g.f. of $N\left(\mu, \frac{\sigma^2}{n}\right)$. Hence $\bar{X}$ has $N\left(\mu, \frac{\sigma^2}{n}\right)$.

E8) You note that in this situation the moment generating function approach is very useful.

The m.g.f. of Z is given by

$$M_Z(t) = E[e^{tZ}]$$

$$= E\left[\exp\left\{t \sum_{i=1}^{n} C_i X_i\right\}\right]$$

$$= \prod_{i=1}^{n} E\left[\exp\left\{t C_i X_i\right\}\right] \text{ (by the independence of } X_i\text{)}$$

$$= \prod_{i=1}^{n} \exp\left\{t C_i \mu_i + \frac{1}{2} t^2 C_i^2 \sigma_i^2\right\}, -\infty < t < \infty$$

since the m.g.f. of X_i is

$$M_{x_i}(t) = \exp\left\{t \mu_i + \frac{1}{2} t^2 \sigma_i^2\right\}, -\infty < t < \infty$$

Hence

$$M_Z(t) = \exp\left\{t \sum_{i=1}^{n} C_i \mu_i + \frac{1}{2} t^2 \sum_{i=1}^{n} C_i^2 \sigma_i^2\right\}$$

for all $-\infty < t < \infty$.

The function on the right side is the m.g.f. of the normal distribution with the mean $\sum_{i=1}^{n} C_i \mu_i$ and the variance $\sum_{i=1}^{n} C_i^2 \sigma_i^2$. Hence, by the uniqueness property of m.g.f.'s (Theorem 1 from Unit 10), it follows that

$$Z = \sum_{i=1}^{n} C_i X_i \text{ is } N\left(\sum_{i=1}^{n} C_i \mu_i, \sum_{i=1}^{n} C_i^2 \sigma_i^2\right).$$

E9) Probability density function of the Chi–square distribution with n degrees of freedom is given by

$$f(z) = \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{\frac{n}{2}-1} e^{-z/2}, \; z > 0$$

$$= 0, \; z \leq 0.$$

Hence, if Z has χ_n^2– distribution, then

$$M_Z(t) = E[e^{tZ}]$$

$$= \int_0^{\infty} e^{tZ} \frac{1}{2^{n/2}\,\Gamma\left(\frac{n}{2}\right)} z^{n/2-1} e^{-z/2}\, dz$$

$$= \frac{1}{2^{n/2}\,\Gamma(\frac{n}{2})} \int_0^{\infty} z^{n/2-1} e^{-z/2(1-2t)}\, dz$$

and the last integral is finite only if $t < 1/2$ by the properties of gamma function. Apply the transformation $z(1-2t) = u$. Then

$$M_z(t) = \frac{1}{2^{n/2}\,\Gamma(\frac{n}{2})} \int_0^{\infty} \left(\frac{u}{1-2t}\right)^{\frac{n}{2}-1} e^{-u/2} \frac{du}{1-2t}$$

$$= \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \frac{1}{(1-2t)^{n/2}} \int_0^{\infty} u^{(n/2)-1} e^{-u/2}\, du$$

$$= \frac{1}{(1-2t)^{n/2}}$$

since

$$\int_0^\infty u^{(n/2)-1} e^{-u/2}\, du = 2^{n/2}\,\Gamma(n/2).$$

from the properties of gamma function (see Sec. 11.3, Unit 11).

E10) $M_{Z_1+Z_2}(t) = E[\exp\{t(Z_1+Z_2)\}]$

$= E[e^{tZ_1}]\, E[e^{tZ_2}]$

$= M_{Z_1}(t)\, M_{Z_2}(t)$

$$= \frac{1}{(1-2t)^{n/2}} \cdot \frac{1}{(1-2t)^{m/2}} = \frac{1}{(1-2t)^{(m+n)/2}}$$

for $t < 1/2$. Since the function $M_{Z_1+Z_2}(t)$ agrees with the m.g.f. of a χ^2_{n+m} – distributed random variable for every $t < 1/2$, in a neighbourhood of zero, it follows that $Z_1 + Z_2$ is χ^2_{n+m}

E11) Check that

$$E[e^{2\lambda n \bar{X} t}] = \frac{1}{(1-2t)^n},\ t < 1/2$$

and use the fact that $\frac{1}{(1-2t)^n}$ is the m.g.f. of χ^2_{2n} –distribution.

E12) Let $Z_i = F_i(X_i)$, $1 \le i \le n$. Then Z_i $1 \le i \le n$ are i.i.d. uniform on [0,1]. Now

$$M_Y(t) = E\left[e^{-2t \sum_{i=1}^{n} \log Z_i} \right]$$

$$= \left(E\left[e^{2t \log z_i} \right] \right)^n$$

$$= \left(E\left[Z_i^{-2t} \right] \right)^n$$

$$= \left(\int_0^1 z_i^{-2t}\, dz \right)^n$$

$$= \frac{1}{(1-2t)^n} \text{ for } t < 1/2$$

which is the m.g.f. of χ^2_{2n} –distribution.

E13) It is clear that $E(U) = 0$. Hence

$$\text{Var}(U) = E[U^2] = \int_0^\infty x^2 \frac{\Gamma\left(\frac{n+1}{2}\right)}{(n\pi)^{1/2}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)} dx$$

$$= 2\int_0^2 \frac{\Gamma\left(\frac{n+1}{2}\right)}{(n\pi)^{1/2}\,\Gamma\left(\frac{n}{2}\right)} \frac{x^2}{\left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}} dx.$$

Apply the transformation $z = \frac{x^2/n}{1+x^2/n}$ and verify that $\text{Var}(U) = \frac{n}{n-2}$ using the properties of the beta density function.

E14) From the table of F–distribution with $n = 5$ we get that $P[U \le 3.365] = .99$ and $P[U \le 1.476] = .90$

$\therefore$ $P[1.476 \le X \le 3.365] = .09$.

E15) Since U has a t– distribution with n degrees of freedom, U can be represented as

$$U = \frac{Y}{\left(\frac{W}{n}\right)^{1/2}}$$

where Y is N(0, 1), W is χ^2_n and Y, W independent.

Hence

$$Z = \frac{Y^2}{W/n}$$

Note that Y^2 is χ^2_1, Y^2 and W are independent and W is χ^2_n. Hence Z has F– distribution with 1 and n degrees of freedom.

E16) If $W = \frac{Y/m}{Z/n}$

where Y is χ^2_m, Z is χ^2_n, Y, Z independent, then $1/W = \frac{Z/n}{Y/n}$

where Z is χ^2_n, Y is χ^2_m and Z, Y independent. Hence 1/W has an F– distribution with n and m degrees of freedom.

E17) a = 3.95, b = 3.14.

APPENDIX

Table of χ^2 Distribution

distribution with n degrees of freedom, this table gives the value of x such that

.005	.01	.025	.05	.10	.20	.25	.30	.40
.0000	.0002	.0010	.0039	.0158	.0642	.1015	.1484	.2750
.0100	.0201	.0506	.1026	.2107	.4463	.5754	.7133	1.022
.0717	.1148	.2158	.3518	.5844	1.005	1.213	1.424	1.869
.2070	.2971	.4844	.7107	1.064	1.649	1.923	2.195	2.753
.4117	.5543	.8312	1.145	1.610	2.343	2.675	3.000	3.655
.6757	.8721	1.237	1.635	2.204	3.070	3.455	3.828	4.570
.9893	1.239	1.690	2.167	2.833	3.822	4.255	4.671	5.493
1.344	1.647	2.180	2.732	3.490	4.594	5.071	5.527	6.423
1.735	2.088	2.700	3.325	4.168	5.380	5.899	6.393	7.357
2.156	2.588	3.247	3.940	4.865	6.179	6.737	7.267	8.295
2.603	3.053	3.816	4.575	5.578	6.989	7.584	8.148	9.237
3.074	3.571	4.404	5.226	6.304	7.807	8.438	9.034	10.18
3.565	4.107	5.009	5.892	7.042	8.634	9.299	9.926	11.13
4.075	4.660	5.629	6.571	7.790	9.467	10.17	10.82	12.08
4.601	5.229	6.262	7.261	8.547	10.31	11.04	11.72	13.03
5.142	5.812	6.908	7.962	9.312	11.15	11.91	12.62	13.98
5.697	6.408	7.564	8.672	10.09	12.00	12.79	13.53	14.94
6.265	7.015	8.231	9.390	10.86	12.86	13.68	14.43	15.89
6.844	7.633	8.907	10.12	11.65	13.72	14.56	15.35	16.85
7.434	8.260	9.591	10.85	12.44	14.58	15.45	16.27	17.81
8.034	8.897	10.28	11.59	13.24	15.44	16.34	17.18	18.77
8.643	9.546	10.98	12.34	14.04	16.31	17.24	18.10	19.73
9.260	10.20	11.69	13.09	14.85	17.19	18.14	19.02	20.69
9.886	10.86	12.40	13.85	15.66	18.06	19.04	19.94	21.65
10.52	11.52	13.12	14.61	16.47	18.94	19.94	20.87	22.62
13.79	14.95	16.79	18.49	20.60	23.36	24.48	25.51	27.44
20.71	22.16	24.43	26.51	29.05	32.34	33.66	34.87	36.16
27.99	29.71	32.36	34.76	37.69	41.45	42.94	44.31	46.86
35.53	37.48	40.48	43.19	46.46	50.64	52.29	53.81	56.62
43.27	45.44	48.76	51.74	55.33	59.90	61.70	63.35	66.40
51.17	53.54	57.15	60.39	64.28	69.21	71.14	72.92	76.19
59.20	61.75	65.65	69.13	73.29	78.56	80.62	82.51	85.99
67.33	70.06	74.22	77.93	82.86	87.95	90.13	92.13	95.81

permission from *Biometrika Tables for Statisticians*, Vol. 1 3rd ed. Cambridge ress, 1966, edited by E. S. Pearson and H.O. Hartley: and from "A new table of oint of the chi-square distribution." *Biometrika*, Vol. 51(1964), pp. 231 - 239, by H.L. space Research Laboratories.

Table of χ^2 Distribution (*Continued*)

.50	.60	.70	.75	.80	.90	.95	.975	.99	.995
.4549	.7083	1.074	1.323	1.642	2.706	3.841	5.024	6.635	7.879
1.386	1.833	2.408	2.773	3.219	4.605	5.991	7.378	9.210	10.60
2.366	2.946	3.665	4.108	4.642	6.251	7.815	9.348	11.34	12.84
3.357	4.045	4.878	5.385	5.989	7.779	9.488	11.14	13.28	14.86
4.351	5.132	6.064	6.626	7.289	9.236	11.07	12.83	15.09	16.75
5.348	6.211	7.231	7.841	8.558	10.64	12.59	14.45	16.81	18.55
6.346	7.283	8.383	9.037	9.803	12.02	14.07	16.01	18.48	20.28
7.344	8.351	9.524	10.22	11.03	13.36	15.51	17.53	20.09	21.95
8.343	9.414	10.66	11.39	12.24	14.68	16.92	19.02	21.67	23.59
9.342	10.47	11.78	12.55	13.44	15.99	18.31	20.48	23.21	25.19
10.34	11.53	12.90	13.70	14.63	17.27	19.68	21.92	24.72	26.76
11.34	12.58	14.01	14.85	15.81	18.55	21.03	23.34	26.22	28.30
12.34	13.64	15.12	15.98	16.98	19.81	22.36	24.74	27.69	29.82
13.34	14.69	16.22	17.12	18.15	21.06	23.68	26.12	29.14	31.32
14.34	15.73	17.32	18.25	19.31	22.31	25.00	27.49	30.58	32.80
15.34	16.78	18.42	19.37	20.47	23.54	26.30	28.85	32.00	34.27
16.34	17.82	19.51	20.49	21.61	24.77	27.59	30.19	33.41	35.72
17.34	18.87	20.60	21.60	22.76	25.99	28.87	31.53	34.81	37.16
18.34	19.91	21.69	22.72	23.90	27.20	30.14	32.85	36.19	38.58
19.34	20.95	22.77	23.83	25.04	28.41	31.41	34.17	37.57	40.00
20.34	21.99	23.86	24.93	26.17	29.62	32.67	35.48	38.93	41.40
21.34	23.03	24.94	26.04	27.30	30.81	33.92	36.78	40.29	42.80
22.34	24.07	26.02	27.14	28.43	32.01	35.17	38.08	41.64	44.18
23.34	25.11	27.10	28.24	29.55	33.20	36.42	39.36	42.98	45.56
24.34	26.14	28.17	29.34	30.68	34.38	37.65	40.65	44.31	46.93
29.34	31.32	33.53	34.80	36.25	40.26	43.77	46.98	50.89	53.67
39.34	41.62	44.16	45.62	47.27	51.81	55.76	59.34	63.69	66.77
49.33	51.89	54.72	56.33	58.16	63.17	67.51	71.42	76.15	79.49
59.33	62.13	65.23	66.98	68.97	74.40	79.08	83.30	88.38	91.95
69.33	72.36	75.69	77.58	79.71	85.53	90.53	95.02	100.4	104.2
79.33	82.57	86.12	88.13	90.41	96.58	101.9	106.6	112.3	116.3
89.33	92.76	96.52	98.65	101.1	107.6	113.1	118.1	124.1	128.3
99.33	102.9	106.9	109.1	111.7	118.5	124.3	129.6	135.8	140.2

Table of the *t* distribution

If X has a t distribution with n degrees of freedom, the table gives the value of x such that $P(X \leq x) = p$.

n	p = 55	.60	.65	.70	.75	.80	.85	.90	.95	.975	.99	.995
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.474	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750
40	.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704
60	.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660
120	.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576

This table is taken from Table III of Fisher & Yates : *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd. London (previously published by Oliver and Boyd Ltd., Edinburgh) and by permission of the authors and publishers.

Table of the 0.95 Quantile of the *F* Distribution

If X has an F distribution with $n1$ and n degrees of freedom the table gives the value of v such that $P_r(N \leq x) = 0.975$.

n \ n_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	248.0	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	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	5.86	5.80	5.75	5.72	5.69	5.66	5.63
5	6.61	5.69	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	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	3.51	3.44	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	3.22	3.15	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	3.01	2.94	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	2.85	2.77	2.70	2.66	2.62	2.58	2.54
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.20	2.16	2.11	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.99	1.95	1.90	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.26	2.21	2.16	2.01	1.93	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	1.92	1.84	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.36	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.75	1.66	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.36	2.21	2.10	2.01	1.94	1.88	1.83	1.67	1.57	1.46	1.39	1.32	1.22	1.00

Adapted with permission from *Biometrika Tables for Statisticians, Vol.* 1, 3rd ed., Cambridge University Press, 1966, edited by E.S. Pearson and H.O. Hartley

If X has an F distribution with m and n degrees of freedom, the table gives the value of x such that $P_r(X \leq x) = 0.975$

n / m	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	984.9	993.1	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	6.52	5.46	5.27	5.17	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.68	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.26	3.20	3.14	3.08
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.59	2.52	2.246	2.40
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.29	2.22	2.16	2.09
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.13	2.07	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	1.94	1.82	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.83	1.71	1.57	1.48	1.39	1.27	1.00

Adapted with permission from *Biometrika Tables for Statisticians, Vol.* 1, 3rd ed., Cambridge University Press, 1966, edited by E.S. Pearson and H.O. Hartley.

UNIT 14 LIMIT THEOREMS

Structure

14.1 INTRODUCTION

In Unit 13, we have discussed different methods for obtaining distribution functions of random variables or random vectors. Even though it is possible to derive these distributions explicity in closed form in some special situations, in general, this is not the case. Computation of the probabilities, even when the probability distribution functions are known, is cumbersome at times. For instance, it is easy to write down the exact probabilities for a binomial distribution with parameters n = 1000 and p = $\frac{1}{50}$. However computing the individual probabilities involve factorials for integers of large order which are impossible to handle even with speed computing facilities.

In this unit, we discuss limit theorems which describe the behaviour of some distributions when the sample size n is large. The limiting distributions can be used for computation of the probabilities approximately.

Chebyshev's inequality is discussed in Section 14.2 and, as an application, weak law of large numbers is derived (which describes the behaviour of the sample mean as n increases). In Sec. 14.3 Binomial distribution with parameters n and p is shown to be approximable by a Poisson distribution whenever n is large and p is such that np is a constant $\lambda > 0$. An important limit theorem, known as the central limit theorem, is studied in Section 14.4. Central limit theorem essentially states that whatever the original distribution is (as long as it has finite variance), the sample mean computed from the observations following that distribution has an approximate normal distribution as long as the sample size (number of observations) is large. An important special case of this result is that binomial distribution can be approximated by an appropriate normal distribution for large samples. This is discussed in Section 14.4. Some examples are presented.

Objectives

After reading this unit, you should be able to

- apply chebyshev's inequality;
- explain the weak law of large numbers;
- apply Poisson or normal approximation to binomial distribution under appropriate conditions; and
- apply the central limit theorem.

14.2 CHEBYSHEV'S INEQUALITY

We prove in this section an important result known as **Chebyshev's inequality.** This inequality is due to the nineteenth century Russian mathematician **P.L. Chebyshev.**

We shall begin with a theorem.

Theorem 1 : Suppose X is a random variable with mean μ and finite variance σ^2. Then for every $\varepsilon > 0$:

$$P\left[\,|X - \mu| \geq \varepsilon\right] \leq \frac{\sigma^2}{\varepsilon^2}. \qquad \text{.....(1)}$$

Proof : We shall prove the theorem for continuous r.vs. The proof in the discrete case is very similar.

Suppose X is a random variable with probability density function f. From the definition of the variance of X, we have

$$\sigma = E\left[(x - \mu)^2\right] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx.$$

Suppose $\varepsilon > 0$ is given. Put $\varepsilon_1 = \frac{\varepsilon}{\sigma}$. Now we divide the integral into three parts as shown in Fig. 1.

$$\sigma^2 = \int_{-\infty}^{\mu - \varepsilon_1\sigma} (x - \mu)^2 f(x)dx + \int_{\mu - \varepsilon_1\sigma}^{\mu + \varepsilon_1\sigma} (x - \mu)^2 f(x)dx + \int_{\mu + \varepsilon_1\sigma}^{\infty} (x - \mu)^2 f(x)dx \qquad \text{.....(2)}$$

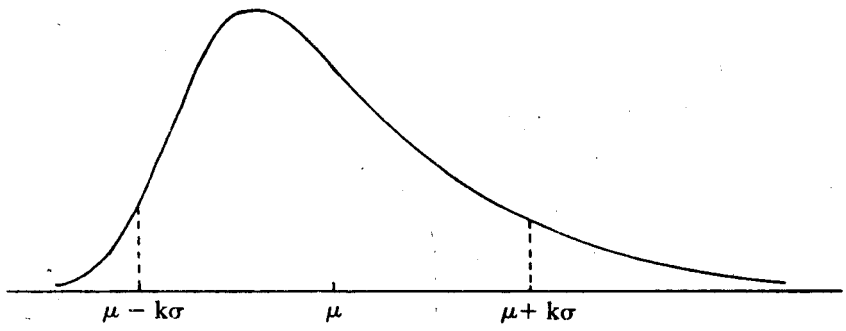


Fig. 1

Since the integrand $(x - \mu)^2 f(x)$ is non–negative, from (2) we get the inequality

$$\sigma^2 \geq \int_{-\infty}^{\mu - \varepsilon_1\sigma} (x - \mu)^2 f(x)dx + \int_{\mu + \varepsilon_1\sigma}^{\infty} (x - \mu)^2 f(x)dx \qquad \text{.....(3)}$$

Now for any $x \in]-\infty, \mu - \varepsilon_1\sigma]$, we have $x \leq \mu - \varepsilon_1\sigma$ which implies that $(x - \mu)^2 \geq \varepsilon^2\sigma^2$. Therefore we get

$$\int_{-\infty}^{\mu - \varepsilon_1\sigma} (x - \mu)^2 f(x)dx \geq \int_{-\infty}^{\mu - \varepsilon_1\sigma} \varepsilon_1^2\sigma^2 f(x)dx$$

$$= \varepsilon_1^2\sigma^2 \int_{-\infty}^{\mu - \varepsilon_1\sigma} f(x)dx.$$

Similarly for $x \in]\mu + \varepsilon_1\sigma, \infty[$ also we have $(x-\mu)^2 \geq \varepsilon_1^2\sigma^2$ and therefore

$$\int_{\mu+\varepsilon_1\sigma}^{\infty} (x-\mu)^2 f(x)dx \geq \varepsilon_1^2\sigma^2 \int_{\mu+\varepsilon_1\sigma}^{\infty} f(x)dx$$

Then by (3) we get

$$\sigma^2 \geq \varepsilon_1^2\sigma^2 \left[\int_{-\infty}^{\mu-\varepsilon_1\sigma} f(x)dx + \int_{\mu+\varepsilon_1\sigma}^{\infty} f(x)dx\right]$$

i.e.
$$\frac{1}{\varepsilon_1^2} \geq \int_{-\infty}^{\mu-\varepsilon_1\sigma} f(x)dx + \int_{\mu+\varepsilon_1\sigma}^{\infty} f(x)dx$$

whenever $\sigma^2 \neq 0$.

Now, by applying Property (iii) of the density function given in Sec. 11.3, Unit 10, we get

$$\frac{1}{\varepsilon_1^2} \geq P[X \leq \mu - \varepsilon_1\sigma] + P[X \geq \mu + \varepsilon\sigma]$$

$$= P[X - \mu \leq -\varepsilon_1\sigma] + P[X - \mu \geq \varepsilon_1\sigma]$$

$$= P[|X - \mu| \geq \varepsilon_1\sigma]$$

That is, $$P[|X - \mu| \geq \varepsilon_1\sigma] \leq \frac{1}{\varepsilon_1^2}. \qquad \text{.....(4)}$$

Substituting $\varepsilon_1 = \frac{\varepsilon}{\sigma}$ in (4), we get the inequality

$$\left[P[|X - \mu| \geq \varepsilon\right] \leq \frac{\sigma^2}{\varepsilon^2}.$$

Chebyshev's inequality also holds when the distribution of X is neither (absolutely) continuous nor discrete. We will not discuss this general case here.

Now we shall make a remark.

Remark 1 : The above result is very general indeed. We need to know nothing about the probability distribution of the random variable X. It could be binomial, normal, beta or gamma or any other distribution. The only restriction is that it should have finite variance. In other words the upper bound is universal in nature. The price we pay for such generality is that the upper bound is not sharp in general. If we know more about the distribution of X, then it might be possible to get a better bound. We shall illustrate this point in the following example.

Example 1 : Suppose X is $N(\mu, \sigma^2)$. Then $E(X) = \mu$ and $Var(X) = \sigma^2$. Let us compute $P[|X - \mu| \geq 2\sigma]$.

Here $\varepsilon = 2\sigma$. By applying Chebychev's inequality we get

$$P\left[|X - \mu| \geq 2\sigma\right] \leq \frac{\sigma^2}{4\sigma^2} = \frac{1}{4} = .25$$

Since we know that the distribution of X is normal, we can directly compute the probability. Then we have

$$P\left(|X - \mu| \geq 2\sigma\right) = P\left[\left|\frac{X - \mu}{\sigma}\right| \geq 2\right].$$

Since $\frac{X - \mu}{\sigma}$ has N(0, 1) as its distribution, from the normal distribution table given in the appendix of Unit 11, we get

$$P\left(\left|\frac{X - \mu}{\sigma}\right| \geq 2\right) = 0.456$$

which is substantially small as compared to the exact value 0.25. Thus in this case we could get a better upperbound by directly using the distribution.

Let us consider another example.

Example 2 : Suppose X is a random variable such that P[X = 1] = 1/2 = P[X = –1]. Let us compute an upper bound for P[|X – μ| > σ].

You can check that E(X) = 0 and Var(X) = 1. Hence, by Chebyshev's inequality, we get that

$$P\left(|X - \mu| > \sigma \right) \leq \frac{\sigma^2}{\sigma^2} = 1.$$

on the other hand, direct calculations show that

$$P\left[|X - \mu| > \sigma \right] = P[|X| \geq 1] = 1.$$

In this example, the upper bound obtained from Chebyshev's inequality as well as the one obtained from using the distribution of X are one and the same.

In the first example you can see an application of Chebyshev's inequality.

Example 3: Suppose a person makes 100 check transactions during a certain period. In balancing his or her check book transactions, suppose he or she rounds off the check entries to the nearest rupee instead of subtracting the exact amount he or she has used. Let us find an upper bound to the probability that the total error he or she has committed exceeds Rs. 5 after 100 transactions.

Let X_i denote the round off error in rupees made for the ith transaction. Then the total error is $X_1 + X_2 + ... + X_{100}$. We can assume that X_i, $1 \leq i \leq 100$ are independent and identically distributed random variables and that each X_i has uniform distribution on $\left[-\frac{1}{2}, \frac{1}{2}\right]$. We are interested in finding an upper bound for the $P\left[|S_{100}| > 5 \right]$ where $S_{100} = X_1 + ... + X_{100}$.

In general, it is difficult and computationally complex to find the exact distribution. However, we can use Chebyshev's inequality to get an upper bound. It is clear that

$$E(S_{100}) = 100\, E(X_1) = 0$$

and

$$Var(S_{100}) = 100\, Var\,(X_1) = \frac{100}{12}.$$

since $E(X_1) = 0$ and $Var(X_1) = \frac{1}{12}$. Therefore by Chebyshev's inequality,

$$P(|S_{100} - 0| > 5) \leq \frac{Var(S_{100})}{25}$$

$$= \frac{100}{12 \times 25}$$

$$= \frac{1}{3}.$$

Here are some exercises for you.

E1) If X is a random variable with $E(X) = \mu$ and $Var(X) = \sigma^2$, find an upper bound for $P[|X - \mu| \geq 3\sigma]$.

E2) Suppose X is a random variable with the exponential density

$$f(x) = e^{-x} \quad \text{for } x > 0$$
$$= 0 \quad \text{for } x \leq 0$$

Let the mean be μ and variance be σ^2 for X. (a) Compute the $P(|X - \mu| \geq 2\sigma)$ using Chebyshev's inequality and (b) compare it with the exact probability obtained from the distribution of X.

E3) If X is a random variable with $E(X) = 3$ and $E(X^2) = 13$, find a lower bound for the probability $P[-2 < X < 8]$.

The above examples and exercises must have given you enough practise to apply Chebyshev's inequality. Now we shall use this inequality to establish an important result.

Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables having mean μ and variance σ^2. We define

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then $\overline{X}_n$ has mean μ and variance $\frac{\sigma^2}{n}$. Hence, by the Chebyshev's inequality, we get

$$P\left[|\overline{X}_n - \mu| \geq \varepsilon \right] \leq \frac{\sigma^2}{n\varepsilon^2}$$

for any $\varepsilon > 0$. If $n \to 0$, then $\frac{\sigma^2}{n\varepsilon^2} \to 0$ and therefore

$$P\left(|\overline{X}_n - \mu| \geq \varepsilon \right) \to 0.$$

In other words, as n grows large, the probability that $\overline{X}_n$ differs from μ by more than any given positive number ε, becomes small. An alternate way of stating this result is as follows :

For any $\varepsilon > 0$, given any positive number δ, we can choose sufficiently large n such that

$$P\left(|\overline{X}_n - \mu| \geq \varepsilon \right) \leq \delta.$$

This result is known as the **weak law of large numbers.** We now state it as a theorem.

Theorem 2 (Weak law of large numbers) : Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and finite variance σ^2.

Let

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then

$$P\left[|\overline{X}_n - \mu| \geq \varepsilon \right] \to 0 \text{ as } n \to \infty.$$

for any $\varepsilon > 0$.

The above theorem is true even when the variance is infinite but the mean μ is finite However this result does not follow as an application of the Chebyshev's inequality in this general set up. The proof in the general case is beyond the scope of this course.

We make a remark here.

Remark 2 : The above theorem only says that the probability that the value of the difference $|\overline{X}_n - X|$ exceeds any fixed number ε, gets smaller and smaller for successively large values of n. The theorem does not say anything about the limiting

case of the actual difference. In fact there is another strong result which talks about the limiting case of the actual values of the differences. This is the reason why Theorem 2 is called 'weak law'. We have not included the stronger result here since it is beyond the level of this course.

Let us see an example.

Example 4 : Suppose a random experiment has two possible outcomes called success (S) and Failure (F). Let p be the probability of a success. Suppose the experiment is repeated independently n times. Let X_i take the value 1 or 0 according as the outcome in the i–th trial of the experiment is a success or a failure. Let us apply Theorem 2 to the set $\{X_i\}_{i=1}^{n}$.

We first note that

$$P[X_i = 1] = p \text{ and } P[X_i = 0] = 1 - p = q,$$

for $1 \leq i \leq n$. Also you can check that $E(X_i) = p$ and $\text{var}(X_i) = p\,q$ for $i = 1,.....n$.

Since the mean and the variance are finite, we can apply the weak law of large numbers for the sequence $\{X_i : 1 \leq i \leq n\}$. Then we have

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \to 0 \text{ as } n \to \infty$$

for every $\varepsilon > 0$ where $S_n = X_1 + X_2 + ... + X_n$. Now, what is $\frac{S_n}{n}$? S_n is the number of successes observed in n trials and therefore $\frac{S_n}{n}$ is the proportion of successes in n trials. Then the above result says that as the number of trials increases, the proportion of successes tends stabilize to the probability of a success. Of course, one of the basic assumptions behind this interpretation is that the random experiment can be repeated.

In the next section we shall discuss another limit theorem which gives an approximation to the binomial distribution.

14.3 POISSON APPROXIMATION TO BINOMIAL DISTRIBUTION

Suppose X is a random variable with the binomial distribution with parameters n and p. Here n is the number of trials and p is the probability of success. From the properties of the binomial distribution, are studied in Unit 7, Block 2, you know that

$$P[X = r] = \binom{n}{r} p^r (1-p)^{n-r}, \quad r = 0, 1, ..., n.$$

Computation of these probabilities when n is large is complicated due to the fact

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

involves n! and (n – r)! which increase rapidly as n increases. You have seen in Unit 7 that

$$E(X) = n\,p \text{ and } \text{Var}(X) = n\,p\,(1-p).$$

Let us look at the limit of the binomial probability

$$\binom{n}{r} p^r (1-p)^{n-r}$$

as $n \to \infty$ such that $np = \lambda$ where $\lambda > 0$ is fixed. You note that as n increases, p has to decrease so that np becomes the constant λ. That is if $n \to \infty$ such that $np = \lambda$, then

$p \to 0$. In other words we are considering a situation where n is "large" and p is "small" and $np = \lambda$. Now

$$\binom{n}{r} p^r (1-p)^{n-r} = \binom{n}{r} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r}$$

$$= \frac{n!}{r!\,(n-r)!} \lambda^r \frac{1}{n^r} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-r}$$

$$= \frac{\lambda^r}{r!} \frac{n!}{(n-r)!} \frac{1}{(n-\lambda)^r} \left(1 - \frac{\lambda}{n}\right)^n.$$

Let us consider the limit of this quantity as $n \to \infty$. It is known from elementary calculus that

$$\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Besides,

$$\frac{n!}{(n-r)!\,(n-\lambda)^r} = \frac{n(n-1).....(n-r+1)}{(n-\lambda)\,(n-\lambda)....(n-\lambda)}$$

which tends to 1 as $n \to \infty$. Thus we have

$$\binom{n}{r} p^r (1-p)^{n-r} \to e^{-\lambda} \frac{\lambda^r}{r!} \text{ as } n \to \infty.$$

such that $np = \lambda$. We summarise the above discussion in the following theorem.

Theorem 3 : When n is large and p is close to 0, the value of $\binom{n}{r} p^r (1-p)^{n-r}$ for the probability $X = r$ of a binomially distributed r.v. X can be approximated by the value $\frac{e^{-\lambda}\lambda^r}{r!}$, $r = 0, 1, 2....$, for probability $Y = r$ where Y is the poisson distributed r.v with mean $\lambda = np$.

The above theorem says that the binomial probability for r successes out of n trials can be approximated by the corresponding Poisson probability for r events to occur with $\lambda = np$ as the mean. This approximation is good whenever n is "large" and p is "small". For the above result to hold, it is not necessary that np is exactly equal to λ. It is sufficient if n and p vary such that $np \to \lambda$ as $n \to \infty$.

In order to get an idea what we mean by n is "large" and p is "small", let us consider the case when $\lambda = 1$. Since $\lambda = np$, we have $p = 1/n$.

For illustrative purposes, we consider the cases $n = 5$, $p = 1/5$ and $n = 100$, $p = 1/100$. See the tables given below.

Table 1

$n = 5, P = \frac{1}{5}, \lambda = 1$		
r	Binomial (n,p)	Poisson (λ)
0	0.328	0.368
1	0.410	0.368
2	0.205	0.184
3	0.051	0.061
4	0.006	0.015
5	0.000	0.003
6	0	0.001

Table 2

$n = 100,\ P = \frac{1}{100},\ \lambda = 1$		
r	Binomial (n,p)	Poisson (λ)
0	0.366032	0.367879
1	0.369730	0.367879
2	0.184865	0.183940
3	0.060999	0.061313
4	0.014942	0.015328
5	0.002898	0.003066
6	0.000463	0.000511
7	0.000063	0.000073
8	0.000007	0.000009
9	0.000001	0.000001
10	0.000000	0.000001

In Tables 1 and 2 given above, the entries are rounded off to the third decimal place in Table 1 and to the sixth decimal place in Table 2. The entries corresponding to r greater than 10 in Table 2 are zero when rounded off and hence are not presented. As can be seen from Tables 1 and 2, the approximation of Binomial by Poisson is not very good when n is small, but the approximation is very good for large n. The agreement is evident, even up to the third decimal place, for every value of r from Table 2.

Let us see an example.

Example 5: Suppose the probability that an item produced by a company is defective is 0.1. Let us compute the probability that a random sample of 10 items of the same kind produced by the same company contains at most one defective item.

If X denotes the number of defective items, then X has the binomial distribution with parameters n= 10 and p = 0.1. We are looking for $P[X \leq 1]$. The exact probability is given by

$$\begin{aligned} P[X \leq 1] &= P[X = 0] + P[X = 1] \\ &= \binom{10}{0}(0.1)^0 (0.9)^{10} + \binom{10}{1}\left((0.1)^1\right)(0.9)^9 \\ &\simeq 0.7361 \end{aligned}$$

On the other hand suppose we approximate the distribution of X by Poisson distribution with $\lambda = np = (10))\ (0.1) = 1$. Then

$$P[X \leq 1] \simeq P[Y \leq 1] = e^{-1} + e^{-1} = 0.7358$$

where Y has the Poisson distribution with mean $\lambda = 1$.

Note the close approximation between the exact probability and the approximate probability.

Try these exercises now.

E4) In a large population, the proportion of people having a certain diseases is 0.01. Find the probability that in a random group of 200 people at least four will have the disease.

E5) Defects in a particular kind of metal sheet occur at an average rate of one per 100 sq. mtr. Find the probability of two or more defects in a sheet of size 5 × 8 sq. mtr.

Thus in this section we have studied that we can approximate a binomial distribution using poisson distribution. In the next section we shall introduce you to another

important limit theorem which gives an approximation to not only binomial distribution but to many other standard distributions.

14.4 CENTRAL LIMIT THEOREM

The central limit theorem (CLT) is one of the most important and useful results in probability theory. We have already seen that the sum of a finite number of independent normal random variables is normally distributed. However the sum of a finite number of independent non-normal random variables need not be normally distributed. Even then, according to the central limit theorem, the sum of a **large** number of independent random variables has a distribution that is approximately normal under general conditions. The CLT provides a simple method of computing the probabilities for the sum of independent random variables approximately. This theorem also suggests the reasoning behind why most of the data observed in practice leads to bell-shaped curves.

Let us now state the main theorem.

Theorem 3 (Central Limit Theorem) : Let $X_1, X_2,$.....be an infinite sequence of independent and identically distributed random variables with mean μ and finite variance σ^2. Then, for any real x,

$$P\left[\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \le x\right] \to \phi(x) \text{ as } n \to \infty \qquad ...(5)$$

where $\phi(x)$ is the standard normal distribution function.

We have omitted the proof because proof of this result involves complex analysis and other concepts which are beyond the scope of this course. Let us try to understand the above statement more clearly. Let $S_n = X_1 + X_2 + ... + X_n$. Then we know that $P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right]$ represents the distribution of the random variable $\frac{S_n - n\mu}{\sigma\sqrt{n}}$. Then the theorem says that the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ is approximately a standard normal distribution for sufficiently large n. Therefore the distribution of S_n will be approximately normal with mean $n\mu$ and variance $n\sigma^2$. In other words the theorem asserts that if $X_1 + X_2 ... + X_n$ are i. i. d. r. v' s of any kind (discrete or continuous) with finite variances, the $\Sigma S_n = X_1 + X_2 + ... + X_n$ will approximately be a normal distribution for sufficiently large n. The importance of the theorem lies in this fact. This theorem has got many applications. An important application is to a sequence of Bernoulli random variables.

Normal approximation to the binomial distribution

Let X_i, $i \ge 1$ be a sequence of i.i.d. random variables such that

$$P[X_i = 1] = p, P[X_i = 0] = 1 - p$$

where $0 < p < 1$.

Observe that $S_n = X_1 + ... + X_n$ has the binomial distribution with parameters n and p. You can check that $E(X_i) = p$ and $Var(X_i) = p(1-p)$ for any i which is finite and positive. An application of the central limit theorem gives the following result:

For every real x,

$$P\left[\frac{S_n - np}{\sqrt{n}\sqrt{[p(1-p)}} \le x\right] \to \phi(x) \quad \text{as } n \to \infty.$$

In other words, for large n

$$P\left[S_n \le np + x\sqrt{np(1-p)}\right] \simeq \phi(x) \qquad ...(6)$$

where $\simeq$ denotes that the quantities on both sides are approximately equal to each other.

An alternate way of interpreting the above approximation is that a binomial distribution tends to be close to a normal distribution for large n. Let us explain this in more detail.

Suppose S_n has binomial distribution with parameters n and p. Then, for $1 \le r \le n$,

$$P\left[S_n \le r\right] = P\left[\frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{r - np}{\sqrt{np(1-p)}}\right]$$

$$\simeq \phi\left[\frac{r - np}{\sqrt{np(1-p)}}\right]$$

for large n by (2). In general, it is computationally difficult to calculate the exact probability

$$P\left[S_n \le r\right] = \sum_{j=0}^{r} \left[\begin{matrix} n \\ j \end{matrix}\right] p^n (1-p)^{n-j}$$

when n is large. A close approximation to this probability can be obtained by computing

$$\phi\left(\frac{r - np}{\sqrt{np(1-p)}}\right),$$

where ϕ is the standard normal distribution function. It has been found from empirical studies that this approximation is good when $n \ge 30$ and a better approximation is obtained by applying a slight correction, namely,

$$\phi\left[\frac{r + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right].$$

Let us illustrate these results by an example.

Example 6 : The ideal size of a first year class in a college is 150. It is known from an earlier data that on the average only 30% of those accepted for admission will actually attend. Suppose the college admits 450 students. What is the probability that more than 150 first year students attend the college?

Let us denote by S_n the number of students that attend the college when n are admitted. Assuming that all the students take independent decision of either attending or not attending the college, we can suppose that S_n has the binomial distribution with parameters n and p = 0.3. Here n = 450 and we are interested in finding the

$$P\left[S_n \ge 150\right].$$

Note that $E(S_n) = np = (450)(0.3) = 135$ and

$$\text{Var}(S_n) = np(1-p) = (135)(.7).$$

Further more

$$P\left[S_n \ge 150\right] = 1 - P\left[S_n < 150\right]$$

$$\simeq 1 - P\left[S_n \le 149\right]$$

and

$$P\left[S_n \le 149\right] \simeq \phi\left[\frac{149 + \frac{1}{2} - 135}{\sqrt{(135)(.7)}}\right]$$

$$= \phi(1.59).$$

Hence

$$P[S_n \geq 150] = 1 - \phi(1.59)$$
$$= .0559$$

This shows that the probability that more than 150 first year students attend is less than 6%.

Let us now consider a different type of application of the central limit theorem.

Example 7: Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables each N(0,1). Then $X_1^2, X_2^2, \ldots$ is a sequence of i.i.d. random variables each with χ_1^2 -distribution. Note that $E(X_i^2) = 1$ and $Var(X_i^2) = 2$ for any i. Hence by central limit theorem we get

$$P\left[\frac{X_1^2 + \ldots + X_n^2 - n}{\sqrt{2n}} \leq x\right] \to \phi(x) \text{ as } n \to \infty.$$

But $S_n = X_1^2 + \ldots + X_n^2$ has χ_n^2 - distribution. What we have shown just now is that if S_n has χ_n^2 - distribution, then $\frac{S_n - n}{\sqrt{2n}}$ has an approximate standard normal distribution for large n. In other words, for every real x,

$$P\left[\frac{S_n - n}{\sqrt{2n}} \leq x\right] \simeq \phi(x)$$

for large n whenever S_n has χ_n^2 -distribution.

We make a remark now.

Remark 3 : The central limit theorem is central to the distribution theory needed for statistical inferential techniques to be developed in Block 4. You must have noted that the distribution of individual X_i in CLT could be discrete or continuous. The only condition that is imposed is that its variance has to be finite. In general, it is not easy to specify the size of n for a good approximation as it depends on the underlying distribution of $\{X_i\}$. However, it is found in practice that, in most cases, a good approximation is obtained whenever n is greater than or equal to 30.

Why don't you try some exercises now.

E6) If X is binomial with n = 100 and p = 1/2, find an approximation for P [X = 50].

E7) Suppose X is binomial with parameters n and p = 0.55. Determine the smallest n for which

$$P\left[\frac{X}{n} > \frac{1}{2}\right] \geq 0.95$$

approximately.

E8) If 10 fair dice are rolled, find the approximate probability that the sum of the numbers observed is between 30 and 40.

E9) Suppose X is binomial with n = 100 and p = 0.1. Find the approximate value of $P(12 \leq X \leq 14)$ using

a) the normal approximation,

b) the poisson approximation, and

c) the binomial distribution.

We will stop our discussion on limit theorem now, though we shall refer to them off and on in the next block. Let us now do quick review of what we have covered in this unit.

14.5 SUMMARY

In this unit, we have:

1) derived Chebyshev's inequality and obtained the weak law of large numbers as a consequence.
2) obtained Poisson approximation to binomial;
3) discussed the central limit theorem and obtained normal approximation to binomial as an application.

As usual we suggest that you go back to the beginning of the unit and see if you have achieved the objectives. We have given our solutions to the exercises in the unit in the last section. Please go through them too. With this we have come to the end of this block.

14.6 SOLUTIONS/ANSWERS

E1) By Chebyshev's inequality,

$$P\left[|X-\mu| \geq 3\sigma\right] \leq \frac{\sigma^2}{9\sigma^2} = \frac{1}{9}$$

E2) a) First note that $E(X) = 1$ and $Var(X) = 1$. Then by Chebyshev's inequality, we get

$$P\left[|X-1| > 2\right] \leq \frac{1}{4} = 0.25.$$

b) $$P\left[|X-1| > 2\right] = P\left[|X| > 3\right]$$

$$= 1 - P\left[|X| \leq 3\right]$$

$$= 1 - \left[1 - e^{-3}\right] \text{ (see Unit 11, Sec. 11.3)}$$

$$= \frac{1}{e^3}$$

$$= 0.05.$$

E3) $$P\left[-2 < X < 8\right] = P\left[-5 < X - 3 < 5\right]$$

$$= P\left[|X-3| \leq 5\right]$$

$$= 1 - P\left[|X-3| > 5\right]$$

By Chebyshev's inequality, we get

$$P\left[|X-3| > 5\right] \leq \frac{4}{25} = 0.16$$

Note that $Var(X) = 4$.

$$\therefore \quad 1 - P\left[|X-3| > 5\right] \geq 1 - 0.16 = 0.84$$

Hence $$P\left[-2 < X < 8\right] \geq 0.84.$$

E4) Let X denote the number of people having the disease. Then X has the binomial distribution with parameters n = 200 and p = 0.01. If we approximate the distribution of X by Poisson distribution N with $\lambda = np = 200(0.01) = 2$, we get

$$P[X \geq 4] \simeq P[N \geq 4] = 1 - P[N < 4]$$

$$\simeq 0.1428.$$

E5) 0.0616.

E6) According to the central limit theorem the distribution of X will be approximately normal with mean 0.5 and variance

$$100 \times \frac{1}{2} \times \frac{1}{2} = 25. \quad \text{Therefore}$$

$$P\left[X = 50\right] = P\left[49.5 \leq X \leq 50.5\right]$$

$$= P\left[\frac{49}{\sqrt{25}} < \frac{X - 0.5}{\sqrt{25}} < \frac{50}{\sqrt{25}}\right]$$

$$\simeq 0.08$$

E7) $$P\left[X > \frac{n}{2}\right] = P\left[\frac{X - np}{\sqrt{np(1-p)}} > \frac{\frac{n}{2} - np}{\sqrt{np(1-p)}}\right].$$

Here p = 0.55 and by the CLT,

$$P\left[\frac{X - np}{\sqrt{np(1-p)}} > \frac{\frac{n}{2} - np}{\sqrt{np(1-p)}}\right]$$

$$P\left[Z > \frac{\frac{n}{2} - np}{\sqrt{np(1-p)}}\right]$$

where Z is N(0.1). In particular

$$P\left[\frac{X}{n} > \frac{1}{2}\right] \geq 0.95$$

provided,

$$P\left[Z > \frac{\frac{n}{2} - np}{\sqrt{np(1-p)}}\right] \geq 0.95.$$

In other words

$$n = \frac{(1.645)^2 p(1-p)}{\left[\frac{1}{2} - p\right]^2}$$

$$= \frac{(1.645)^2 (0.55)(0.45)}{(0.05)^2}$$

$$= 268.$$

E8) 0.65.

E9) (a) 0.2417 (b) 0.5710 (c) 0.5642

Notes

NOTES

NOTES

UTTAR PRADESH
RAJARSHI TANDON OPEN UNIVERSITY

UGMM - 11

PROBABILITY AND STATISTICS COURSE

Block

4

ELEMENTS OF STATISTICAL INFERENCE

Course Design Committee

Prof. S.K. Mitra *(Chairman)*
Indian Statistical Institute
New Delhi

Prof. A.M. Goon
Presidency College
Calcutta

Prof. J. Medhi
Guwahati

Prof. B.L.S. Prakasa Rao
Indian Statistical Institute
New Delhi

Prof. Aloke Dey
Indian Statistical Institute
New Delhi

Prof. K. Balasubramanian
Indian Statistical Institute
New Delhi

Prof. D.D. Joshi
Ex.Pro.Vice-Chancellor
IGNOU

Dr. V. Madan
School of Sciences
IGNOU

Dr. Poornima Mital
School of Sciences
IGNOU

Dr. Manik Patwardhan
School of Sciences
IGNOU

Dr. Sujatha Varma
School of Sciences
IGNOU

Block Preparation Team

Prof. S. K. Mitra (Editor)
ISI, New Delhi

Prof. Aloke Dey (*Co-editor*)
ISI, New Delhi

Prof. R. K. Bose
School of Sciences
IGNOU

Dr. V. K. Gupta
Indian Agricultural Statistics Research Institute
Delhi

Course Coordinator : Prof. R. K. Bose

Acknowledgement
To Dr. Sujatha Varma for her help in block preparation.

Production

Mr. Balakrishana Selvaraj
Registrar (PPD)
IGNOU

March,1994

ISBN-81-7263-569-9

ELEMENTS OF STATISTICAL INFERENCE

This is the last block which you will be studying for the Probability and Statistics Course. In this block, we shall be dealing with the basic concepts of statistical inferences and also with some methods of making inferences on the basis of sample observations. Since all the procedures described in this block are based on sample observations, you will find a discussion on Random sampling and Sampling distributions in Unit 15. In the same Unit, you will be introduced to the basic concepts regarding point estimation, testing of hypothesis and interval estimation.

Some common methods of point estimation are discussed in Unit 16. In Units 17 and 18, you will find discussion on testing of hypothesis and interval estimation. Here procedures for testing hypothesis involving the parameters of some important distributions and procedures for constructing confidence intervals for parameters are discussed.

In Unit 15-18, we have included a number of illustrative examples. If you go through these examples carefully, you will have a better understanding of the concepts discussed. These will also serve as a guide for solving exercises.

Notations and Symbols

Ω	:	Parameter space
$\hat{\theta}$	:	An estimator of θ, a parameter.
H(or H_0)	:	Null Hypothesis
A(or H_1)	:	Alternative Hypothesis
C	:	Critical Region
C_0	:	Optimum critical region
$L_n(\theta)$	:	$\prod_{i=1}^{n} f(X_i; \theta)$ (i.e. Likelihood function)
$\lambda(X)$	:	$\frac{\sup_{\theta \in \Omega_0} L(\underline{\theta} \mid \underline{X})}{\sup_{\theta \in \Omega} L(\underline{\theta} \mid \underline{X})}$ (i.e. Likelihood Ratio)
$\alpha(\theta)$	:	Probability of Type I error
$\beta(\theta)$	:	Probability of Type II error
$\gamma(\theta)$	:	$1 - \beta(\theta)$
$I_n(\theta)$	:	Fisher information in the sample (X_1 ....., X_n)

Also see lists in previous blocks.

UNIT 15 GENERAL INTRODUCTION

Structure

15.1 INTRODUCTION

In the earlier blocks, you have studied the concepts of probability theory and its various applications in model building for different random phenomena. In probability theory, we proceed from a known population and derive probabilities of events associated with a phenomenon. The basic problems in Statistics, on the other hand, is concerned with the reverse process of drawing conclusions (or, inferences) about a population on the basis of a sample.

In order to be able to make probabilistic statements about a population on the basis of a sample, the sample itself has to be chosen in a appropriate manner. Broadly, the methods or sample selection go by the name of random sampling. In this unit, we confine attention to simple random sampling only. We shall also discuss, in a general way, the problem of statistical inference, viz., that of point estimation and those of testing of hypotheses and interval estimation.

Objectives

After reading of this unit, you should be able to

- derive some basic properties of sample statistics
- define sampling distribution of a statistic and derive some basic sampling distributions
- define the important properties of an estimator
- define the important concepts relating to interval estimation and testing of hypothesis.

15.2 INDUCTIVE INFERENCE

Scientific progress is often ascribed to experimentation. A research worker performs an experiment and obtains some data and on the basis of the data so collected, some conclusions are drawn. The conclusions usually go beyond the materials and operations of the particular experiment, that is, the research worker may generalize from a particular experiment to the class of similar experiments. This type of extension from the particular to the general is called inductive inference and is one of the ways in which new knowledge is acquired.

Inductive inference is known to be a hazardous process because of the uncertainty present in the inference. One simply cannot make perfectly certain generalizations. However, uncertain inferences can be made and the degree of uncertainty measured if the experiment has been performed according to certain well defined principles. One function of Statistics is the provision of techniques for making inductive inferences and for measuring the degree of uncertainty of such inferences. Uncertainty is measured in terms of probability of making a wrong inference.

Let us illustrate inductive inference by an example. Suppose we have a storage bin containing say 1,00,000 seeds of flowers. It is known, that these seeds will grow into plants with either a white or a red flower. The information sought is: how many (or, what proportion) of these seeds will produce plants with red flowers? The only way in which we can be sure that this question is answered correctly is to plant every seed in the bin and count the number producing red flowers. However, this is not feasible as the seeds are for sale. Even if the seeds were not for sale, one would prefer to have an answer without going through such an enormous effort of planting the seeds and see them flower. At this stage, a natural question to ask is: can we plant a few of the seeds and on the basis of the colours of flours observed on individual plants make a statement as to how many of these will produce plants with red flowers? The answer to this question is that we cannot make an exact prediction as to how many plants with red flowers the seeds in the bin would produce, but we can make a probabilistic statement if we select the few seeds in a certain fashion. This is inductive inference: We select a few of the 1,00,000 seeds, plant, observe the number of red flowers and on the basis of these few make a statement as to how many of the 1,00,000 will produce plants with red flowers; from the knowledge of the colour of a few, we generalize to the whole 1,00,000. We cannot be certain of our answer but we can have confidence in our statement in a frequency-ratio probability sense.

15.3 RANDOM SAMPLING

As seen in the previous Section, we will attempt to make inference on the basis of a sample, chosen suitably from the population. We now formalize some concepts regarding sampling.

Definition 1 : The totality of elements which are under discussion and about which information is sought is called the (Target) population.

In the example of the previous Section, 1,00,000 seeds in the bin form the target population. The target population may be (i) the totality of dairy cattle in a state, or, (ii) the prices of a certain commodity on a given day, or, (iii) the collection of all hypothetical sequence of heads and tails obtained by tossing a coin an infinite number of times. The important point is that the target population must be well defined. It could be real as in the case of population in examples (i) and (ii) above or hypothetical, as in the case of example (iii).

Now, consider a statistical experiment that results in outcomes x, which are the values assumed by a random variable X. For instance, in the example of the previous section, if we select one seed from the bin containing 1,00,000 seeds at random, then on planting this seed, the resulting plant will either produce a red flower or a white flower. Here the random variable X takes two values, $X = 1$, if the plant gives rise to a red flower and $X = 0$, if the plant gives rise to a white flower. Suppose the probability of the seed producing a plant with red flower is p and that with a white flower be $q = (1 - p)$, then the random variable X clearly has a Bernoulli distribution with probability of "success" (producing a plant with red flower) p, which may be unknown. Thus, we have a random variable X with known form of distribution function with its parameters p, possibly unknown. Let F be the distribution function (d.f) of X. In practice, F will not be known completely in the sense that one or more parameters associated with F will be unknown, (for

example, p is unknown in the above illustration) and, the statistician desires to make inferences about the unknown parameters. For this purpose, the statistician can obtain n independent observations $x_1, x_2, \ldots, x_n$ assumed by the random variable X. Each x_i can be regarded as the value assumed by the random variable X_i, $i = 1, 2, \ldots, n$, where $X_1, X_2, \ldots, X_n$ are independent random variables with common d.f.F. The observed values $(x_1, x_2, \ldots, x_n)$ are then values assumed by the random variables $(X_1, X_2, \ldots, X_n)$. The set $(X_1, X_2, \ldots, X_n)$ is a sample of size n taken from a population with distribution function F and the set of values $(x_1, x_2, \ldots, x_n)$ is called a realization of the sample.

Definition 2: Let X be a random variable with distribution function F and let $X_1, X_2, \ldots, X_n$ be independently and identically distributed (i.i.d.) random variables with common d.f. F. Then, the collection $X_1, X_2, \ldots, X_n$ is known as a random sample of size n from the d.f. F, or simply, as n **independent observations on the random variable X.**

Definition 3: A statistic is a function of observable random variables which does not contain any unknown parameter. Let $X_1, \ldots, X_n$ be a random sample from X and let $x_1, \ldots, x_n$ be the values assumed by the sample. Let H be a function defined for the n-tuple $(x_1, \ldots, x_n)$. $\hat{Y} = H(X_1, \ldots, X_n)$ is defined to be a statistic, assuming the value $y = H(x_1, \ldots, x_n)$.

Example 1 : Let $X_1, X_2, \ldots, X_n$ be a random sample from a d.f.F. Then $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$, the sample mean, is a statistic. Similarly, let $S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$. Clearly, S^2 is also a statistic.

Another example of a statistic would be the smallest (or, the largest) of the sample observations. It follows that the difference of the largest and smallest observation in the sample, called the sample range, is also a statistic.

Example 2 : Let the random variable X follow a normal distribution with known mean μ and unknown variance σ^2. Then $X - \mu$ is a statistic but $(X - \mu)/\sigma$ is not.

E1) Let X be a random variable that takes only two values 1 and 0 with respective probabilities p and $q = 1 - p$, where p is possibly unknown. A random sample of size 5 is drawn from X and the realizations are 0, 1, 1, 1, 0. Compute $\overline{X}$ and S^2, where these statistics are defined in Example 1.

15.4 SAMPLING DISTRIBUTIONS RELATED TO NORMAL DISTRIBUTION

We now investigate certain distributions that arise in sampling from a normal population. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Let, as before, $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$,

$S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$. Then, the following results are true.

(a) $\overline{X}$ has a normal distribution with mean μ and variance σ^2/n.

(b) $(n-1)S^2/\sigma^2$ has a chi-square distribution with $(n-1)$ degrees of freedom (D.F.)

(c) $\bar{X}$ and S^2 are independently distributed.

You have already seen in Unit 12 (Section 12.5) that if X and Y are two independent random variables such that X has a normal distribution with mean zero and unit variance and Y has a chi-square distribution of n D.F., then the random variable $Z = X/\sqrt{(Y/n)}$ is said to have t-distribution with n D.F. From this fact and the properties (a), (b), (c) above, it follows that the statistic

$$U = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\left\{\frac{(n-1)S^2}{\sigma^2}/(n-1)\right\}^{1/2}} = \frac{\sqrt{n}\,(\bar{X}-\mu)}{\sqrt{S^2}}$$

follows a t-distribution with $(n-1)$ D.F.

Again, recall from Unit 12 (Section 12.6) that if X and Y are independent chi-square random variables with m and n D.F. respectively, then the random variable $(X/m)/(Y/n)$ is said to follow an F-Distribution with (m, n) D.F. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent samples of size m and n respectively from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ and let

$$\bar{X} = m^{-1}\sum_{i=1}^{m} X_i,\ \bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i,\ S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2,\ S_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

It follows then that the statistic (by (b) above).

$$W = \frac{S_1^2}{S_2^2}$$

follows an F-distribution with $(m-1, n-1)$ D.F.

E2) Let $X_1, X_2, \ldots X_n$ be a random sample of size n from $N(0, \sigma^2)$. What is the distribution of $S_0^2 = \sum_{i=1}^{n} X_i^2/\sigma^2$?

E3) If $X_1, X_2, \ldots X_n$ is a random sample from $N(\mu, 1)$, what is the distribution of $T = \sum_{i=1}^{n}(X_i-\bar{X})^2$?

15.5 POINT ESTIMATION

In this section we shall discuss the basics of the theory of point estimation. The problems of point estimation of a parameter can be visualized as follows:

Let X be a random variable with df $F(x;\theta)$ where θ is a parameter. The parameter could be a scalar or vector. The parameter θ is a scalar if it is real-valued and θ is a vector parameter if it is of the form $\theta = (\theta_1, \theta_2 \ldots \theta_k)$, where for $i = 1, 2, \ldots, k$, θ_i is real-valued and k is finite. The set of possible values of θ is called the parameter space Ω. In the discussion that follows, Ω will mostly be a subset of the real line R or R^k, the k-dimensional Euclidean space for some finite k. It is assumed that the functional form of F is known, except for θ. For example, F might be the df of a normal distribution with mean zero and variance σ^2

(unknown), or, F could be the df of a binomial distribution with n, the number of trial known and p, the probability of success, unknown. The first one is an example of a continuous d.f. and the second one of a discrete d.f. Of course, it is possible that F is a mixture of a discrete and a continuous d.f. However, in what follows, we shall restrict our attention to the case where F is the d.f. of either a discrete random variable or a continuous random variable.

Suppose a random sample of *n* observations, $X_1, X_2, ..., X_n$ is taken from a df $F(x ; \theta)$ and suppose $x_1, x_2, ..., x_n$ are the realizations of $X_1, X_2, ..., X_n$ i.e., the observed data is $(x_1, x_2, ..., x_n)$. The problem of point estimation is then to estimate the unknown parameter θ through a suitable statistic, which by definition 3, is a function of $X_1, X_2, ..., X_n$ and is free of θ. If the chosen statistic is $T (X_1, X_2, ..., X_n)$ then $T (X_1, X_2, ..., X_n)$ is called an estimator of θ. If we substitute the actual observations $x_1, x_2, ..., x_n$ in the functional form of $T (X_1, X_2, ..., X_n)$ and compute the value of $T (x_1, x_2, ..., x_n)$, then this value is called an estimate of θ. An estimator of θ is usually written as $\hat{\theta} = T (X_1, \ldots, X_n)$.

Example 3 : Let $X_1, X_2, ..., X_n$ be a random sample from a Poisson distribution with parameter λ, which is unknown, Then, $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ is an estimator of λ.

Similarly the mean of $(n - 2)$ observations, obtained by discarding the smallest and the largest observations in the sample is also an estimator of λ.

Example 4 : Suppose $X_1, X_2, ..., X_n$ is a random sample from a binomial population with n (=number of trials) = 10, and p, the probability of success, unknown.

Then, $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ and $(X_1 + X_n)/2$ are **both estimators** of p.

From the above example, it is clear that we need some criterion to choose among many possible estimators. Below, we give some criteria which can be used to judge the performance ("goodness") of an estimator.

Definition 4 :(Unbiased Estimator). An estimator $T (X_1, X_2, ..., X_n)$ is said to be an unbiased estimator of $g (\theta)$ (a known function of θ) if the expectation of $T (X_1, X_2, ..., X_n)$ exists and is equal to $g (\theta)$ for all θ in the parameter space Ω. In symbols, $T (X_1, X_2, ..., X_n)$ is unbiased for $g (\theta)$ if

$E_\theta \left[T (X_1, X_2, ..., X_n) \right] = g (\theta)$ for all $\theta \in \Omega$ where E_θ denotes the expectation taken when θ is the parameter.

The intuitive implication of the concept of unbiasedness is as follows: Even though the value of T may not be equal to $g (\theta)$ or near to $g (\theta)$ for a particular realization $(x_1, x_2, ..., x_n)$, on the average, T is close to $g (\theta)$. In other words, in repeated sampling of *n* observations from the population, the average of $T (X_1, X_2, ..., X_n)$ is equal to $g (\theta)$. We illustrate this concept through some examples.

Example 5 : Let $X_1, X_2, ..., X_n$ be n independent Bernoulli random variables, each with the same probability p of success. We wish to obtain an unbiased estimator of p based on the observations $X_1, X_2, ..., X_n$. Now, we know that

$$P(X_i = 1) = p \text{ and } P(X_i = 0) = 1 - p = q \text{ (say)}$$

Hence

$$E_p(X_i) = p \text{ for any } i = 1, 2, \ldots, n.$$

Thus, every one of the observations can be considered as an unbiased estimate of p. There are other unbiased estimators of p as well, e.g. $\bar{X} = n^{-1}(X_1 + X_2 + \ldots + X_n)$ is also unbiased for p, because,

$$E_p(\bar{X}) = n^{-1} \sum_{i=1}^{n} E_p(X_i) = p.$$

Example 6 : Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with known mean μ_0 and unknown variance σ^2. So $\theta = \sigma^2$ and $\Omega = \{\theta ; \theta > 0\}$. Consider the estimator $(X_i - \mu_0)^2$. Since $(X_i - \mu_0)/\sigma$ is distributed as normal with mean zero and unit variance, it follows that $\{(X_i - \mu_0)/\sigma\}^2$ is distributed as chi-square with one D.F.

The mean of $\{(X_i - \mu_0)/\sigma\}^2$ is unity and hence

$$E_\theta (X_i - \mu_0)^2 = \sigma^2 = \theta$$

Since the above identity holds for each $i = 1, 2, \ldots, n$, we have

$$E_\theta \sum_{i=1}^{n} (X_i - \mu_0)^2 = \sigma^2 = \theta$$

and hence $\hat{\theta} = n^{-1} \sum_{i=1}^{n} (X_i - \mu_0)^2$ is an unbiased estimator of $\theta = \sigma^2$

Example 7: Now let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Here $\Omega = \{(\mu, \sigma^2) ; -\infty < \mu < \infty ; \sigma^2 > 0\}$. Clearly, in this case, $n^{-1} \sum_{i=1}^{n} (X_i - \mu)^2$ is not a statistic and hence cannot be used as an estimator of σ^2. Let us consider the estimator

$$S_0^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \text{ where } \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$$

We know that $\sum_{i=1}^{n} (X_i - \bar{X})^2/\sigma^2$ has a chi-square distribution with $(n-1)$ D.F. and therefore

$$E_{\sigma^2}\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right] = (n-1)\sigma^2.$$

Hence $E_{\sigma^2}(S_0^2) = \frac{n-1}{n}\sigma^2$

so that S_0^2 is not an unbiased estimator of σ^2. However,

$$S^2 = nS_0^2/(n-1) = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator of σ^2. Note that S^2 is defined only when $n \geq 2$.

Example 8 : Suppose it is desired to estimate the standard deviation, σ, of a normal population with unknown mean μ and unknown variance σ^2. Since

$$S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \text{ is an unbiased}$$

estimator of σ^2 one may be tempted to use S as an unbiased estimator of σ. Using the fact that $(n-1)S^2/\sigma^2$ is distributed as Chi-square with $(n-1)$ D.F., it can be shown that

$$E_\sigma(S) = \frac{\sigma \Gamma(n/2)}{\Gamma\left(\frac{n-1}{2}\right)} \sqrt{\frac{2}{n-1}}$$

so that S is **not** unbiased for σ.

E4) Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Show that $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ is unbiased for λ. Is

$S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ also unbiased for λ ?

E5) Let $X_1, X_2, \ldots, X_n$ be n independent Bernoulli random variables with constant unknown probability of success p. Let $\bar{X} = n^{-1} \sum_{i=1} X_i$. Show that $\bar{X}(1-\bar{X})$ is not an unbiased estimator of $p(1-p)$.

Note that sometimes, an unbiased estimator may turn out to be absurd, as shown in Example 9.

Example 9: Let X_1 be a random sample of just one observation from a Poisson distribution with parameter λ. Suppose it is desired to estimate $g(\lambda) = \exp(-3\lambda)$. Consider the estimator $T = (-2)^{X_1}$.

Then

$$E_\theta(T) = e^{-\lambda} \sum_{X_1=0}^{\infty} (-2)^{X_1} \frac{\lambda^{X_1}}{X_1!} = e^{-\lambda} \sum_{X_1=0}^{\infty} \frac{(-2\lambda)^{X_1}}{X_1!} = e^{-\lambda} \cdot e^{2\lambda} = g(\lambda).$$

Therefore, $T = (-2)^{X_1}$ is an unbiased estimator for $g(\lambda) = e^{-3\lambda}$. However $T > 0$ if X_1 is even and is negative if X_1 is odd, which leads to an absurd situation as $g(\lambda) > 0$.

At this stage a natural question to ask is : how to choose from a collection of unbiased estimators for the same parameter, whenever they exist ? Our discussion is restricted to only those unbiased estimators which have finite variances. Let T be an

unbiased estimator of a scalar parameter θ. Then, $T - \theta$ measures the departure of the estimator T from θ. Since $T - \theta$ is a function of observations $X_1, X_2, ..., X_n$ as well as of θ a measure of deviation of T from θ may be chosen as $E_\theta (T - \theta)$. However, since T is unbiased for θ, this expectation is always zero, so this is not a meaningful measure. A better measure of deviation can be considered as $E_\theta | T - \theta |$ or $E_\theta | T - \theta |^p$, $p \geq 1$. A convenient choice of p is 2 and the precision of T is measured by the quantity $E_\theta | T - \theta |^2$. Since T is unbiased for θ, this measure is simply the variance of T, denoted henceforth as $Var_\theta (T)$. The larger the variance of T, the greater is the departure of T from the true value θ, on an average. On this basis, we may prefer an unbiased estimator T_1 over another unbiased estimator T_2 (both of θ), if

$$Var_\theta (T_1) \leq Var_\theta (T_2) \text{ for all } \theta \in \Omega,$$

with strict inequality of at least one $\theta \in \Omega$.

Consider now the class of all unbiased estimators of θ that have finite variances. Is it possible to find an estimator in this class that has the smallest variance ? If such an estimator exists, it will be called a minimum variance unbiased estimator of θ. We shall discuss in the next unit more about this aspect.

We now turn to another criterion for the choice of an estimator.

Definition 5: (Consistent Estimator). An estimator $T(X_1, X_2, \ldots, X_n)$ is said to be consistent for θ if $T_n = T(X_1, X_2, \ldots, X_n)$ converges in probability to θ, as *n* increases indefinitely, that is T_n a consistent estimator of θ if $P_\theta (|T_n - \theta| > \varepsilon) \to 0$.

as $n \to \infty$ for every $\varepsilon > 0$.

Example 10: Let $X_1, X_2, ..., X_n$ constitute n independent Bernoulli random variables with the same probability p of success. Then, $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ is a consistent estimator of p, because,

$$P_p \left(|\overline{X} - p| > \varepsilon \right) \to 0 \text{ as } n \to \infty \text{ for every } \varepsilon > 0.$$

This can be seen as follows : By Chebychev's inequality

$$P_p \left(|\overline{X} - p| > \varepsilon \right) \leq \frac{1}{\varepsilon^2} Var_p (\overline{X}).$$

Now, we know that $\overline{X}$ has a binomial distribution with parameters n and p and hence

$$Var_p (\overline{X}) = \frac{1}{n^2} Var \left(\sum_{i=1}^{n} X_i \right) = p(1-p)/n \to 0 \text{ as } n \to \infty.$$

Therefore, $\overline{X}$ is a consistent estimator of p.

Example 11: Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from $N(\mu, \sigma^2)$, both μ and σ^2 being unknown. Let $S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

Then,

$$P_{\sigma^2}\left(\left|S^2-\sigma^2\right|>\varepsilon\right)\le\frac{1}{\varepsilon^2}E_{\sigma^2}\left(S^2-\sigma^2\right)^2=\frac{1}{\varepsilon^2}\mathrm{Var}_{\sigma^2}\left(S^2\right)$$

But

$$\mathrm{Var}_{\sigma^2}(S^2)=\frac{1}{(n-1)^2}\mathrm{Var}\left(\Sigma(X_i-\bar{X})^2\right)$$

$$=\frac{\sigma^4}{(n-1)^2}\mathrm{Var}\left(\Sigma(X_i-\bar{X})^2/\sigma^2\right)$$

$$=\frac{2(n-1)\sigma^4}{(n-1)^2}=2\sigma^4/(n-1),$$

where in deriving the above variance, we have used the fact that $\sum_{i=1}^{n}(X_i-\bar{X})^2/\sigma^2$ has a chi-square distribution with $(n-1)$ D.F. and the variance of such a distribution is $2(n-1)$. Thus, $P_{\sigma^2}\left(\left|S^2-\sigma^2\right|>\varepsilon\right)\to 0$ as $n\to\infty$ for every $\varepsilon>0$ and S^2 is a consistent estimator of σ^2.

From the above examples, it is clear that a set of sufficient conditions for T_n to be a consistent estimator of θ is :

(i) $E_\theta(T_n)\to\theta$ as $n\to\infty$

(ii) $\mathrm{Var}_\theta(T_n)\to 0$ as $n\to\infty$.

A consistent estimator for θ need not be unbiased. If there is an unbiased estimator for θ, then condition (i) above is automatically satisfied and only condition (ii) has to be checked for consistency.

E6) Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Show that the sample mean is a consistent estimator of μ.

E7) Let $X_1, X_2, \ldots, X_n$ constitute n independent Bernoulli random variables with constant probability of success, p, which is unknown. Obtain a consistent but biased estimator of p.

15.6 TESTING OF HYPOTHESIS

In this section, you will be introduced to some basic notions about testing of hypothesis. A hypothesis-testing problem results from questions of the type: "Does smoking increase the risk of cancer?" "Does a certain feed increase the milk yield of dairy cattle?" Thus, we have an underlying parameter θ (in our examples, this might be cancer rate, milk yield) and we wish to known whether it changes in specified ways (e.g. does it increase ?) when an element of a system is changed.

To fix ideas, we now formally define the concepts involved. As before, let X be a random variable with d.f $F(x;\theta)$ where $\theta \in \Omega$, the parameter space. We shall assume that the functional form of F is known except for the parameter θ. Also, we assume that Ω has at least two points.

A (parametric) hypothesis is an assertion about the unknown parameter θ.

Example 12: In coin tossing experiments, a question often asked is whether the coin is unbiased (fair), that is, whether the probability of getting heads or tails is the same, 0.5. Thus, in this case, one may set up a hypothesis, $H : p = 1/2$, where p i the unknown probability of obtaining a head in a single toss.

Example 13: A manufacturer of dry cells claims that the cells manufactured by hir last 30 hours. To test this claim, one may set up a hypothesis, $H : \mu = 30$, where μ denotes the average life of a dry cell.

In the above two examples, the hypothesis taken were such that the difference between the unknown parameter and the hypothetical value were zero (null). This gave rise to the term Null Hypothesis. However, the term null hypothesis is not restricted to the hypotheses of the kind described in Examples 12 and 13. We shall refer to any hypothesis under test as the Null hypothesis. Note that a null hypothesis is a statement about the parameter (s) belonging to a subset Ω_0 of Ω. The null hypothesis can therefore be specified as $H : \theta \in \Omega_0$.

Corresponding to any null hypothesis, the statement $A : \theta \in \Omega_1 = \Omega - \Omega_0$ usuall referred to as the **alternative** hypothesis.

Definition 7: If Ω_0 (Ω_1) contains only one point, we say that H (A) is simple; otherwise, we say that the hypothesis (H of A) is composite. Clearly, if a hypothesis is simple, the probability distribution of X is completely specified under the hypothesis.

Example 14: Let X be a random variable having a normal distribution with mean μ and variance σ^2. If both μ and σ^2 are unknown $\Omega = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. The hypotheses, $H : \mu > \mu_0, \sigma^2 > 0$, $H : \mu \leq \mu_0, \sigma^2 > 0$, $H : \mu = \mu_0, \sigma^2 0$, where μ_0 is a specified constant, are all composite. On the other hand, if σ^2 is known, then $H : \mu = \mu_0$ is a simple hypothesis.

The problem of testing of hypothesis can now be described as follows: Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population with d.f. $F(x;\theta)$. Let X^n denote the sample space corresponding to $X_1, X_2, \ldots, X_n$. Choose a subset $C \subset X^n$. We call C a critical region. Suppose $(X_1, X_2, \ldots, X_n)$ is the observed sample. Then, the test procedure is:

reject H if $(X_1, X_2, \ldots, X_n) \in C$

and do not reject H (or, accept H) if $(X_1, X_2, \ldots, X_n) \notin C$.

Such a procedure is called a non-randomized test for testing the null hypothesis H against the alternative A.

Some caution should be exercised when the second of the above two actions is taken. Accepting the hypothesis H does not necessarily mean that we conclude definitely that $\theta \in \Omega_0$; instead what is meant is that on the basis of the data

available there is no evidence for not supporting the hypothesis that $\theta \in \Omega_0$.

From the above discussion it is clear that one might commit two kinds of error with the test procedure described above. The test procedure may lead to rejection of H when really it is true, or it might lead to rejection of A (or, acceptance of H) when in fact A is true. The two types of errors can be represented in a tabular form, as shown below:

		Action Taken	
		H accepted	H rejected
State of	H True	Correct	Type I error
Nature	A True	Type II error	Correct

The problem then is to devise test procedures to control both types of errors. Ideally, one would prefer a test procedure that minimizes α and β where

α = Probability of Type I error

and β = Probability of Type II error

However, this in general is not possible. Let

$$\alpha(\theta) = P_\theta(\text{Reject H}) \text{ and } \beta(\theta) = P_\theta(\text{Accept H})$$

Then $\alpha(\theta)$ denotes the probability of Type I error when $\theta \in \Omega_0$ and $\beta(\theta)$ is the probability of Type II error when $\theta \in \Omega_1$. If $\gamma(\theta) = 1 - \beta(\theta)$, then $\gamma(\theta)$ is called the **power** of the test at $\theta \in \Omega_1$. In constructing a test procedure, we fix the probability of the Type I error to a desired small level and choose a test for which $\gamma(\theta)$ is maximum (equivalently, $\beta(\theta)$ is minimum). Given $0 \le \alpha \le 1$, our interest is then to construct a test procedure for which

$$\alpha(\theta) \le \alpha \text{ for all } \theta \in \Omega_0$$

and $\gamma(\theta) = 1 - \beta(\theta)$ is as large as possible for all $\theta \in \Omega_1$. The number α is called the **level of significance** of the test. In practice, we choose α to be small, usually 0.05 or 0.01. To illustrate these ideas, we consider an example.

Example 15: Let X follow a normal distribution, N (μ, 1), where μ is unknown. Let the problem be to test H : $\mu = 0$ against A : $\mu = 1$. Clearly, both H and A are simple hypotheses. Suppose $X_1, X_2, \ldots, X_n$ is a random sample of size n drawn from N (μ, 1) and let $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$ be the sample mean. Since $\overline{X}$ is an estimator of μ, it is natural to base our test on $\overline{X}$ and we may say that A is true if $\overline{X}$ is large and H is true if $\overline{X}$ is small. Thus, our test procedure may be

reject H if $\overline{X} > k$

accept H if $\overline{X} \le k$

where k is a suitable constant. What are the probabilities of type I and II errors for this test? We have

$$\alpha(\mu) = P_\mu(\overline{X} > k)$$

If we want a test with level of significance α (a given constant) then the constant k

should be chosen to satisfy

$$p(\bar{X} > k \mid H) = \alpha.$$

But, we know that $(\bar{X} - \mu)/(1/\sqrt{n})$ has a N (0, 1) distribution. **Hence, under H,** $(\bar{X} - 0)/(1/\sqrt{n})$ follows a N (0, 1) distribution. Therefore

$$\alpha = P(\bar{X} > k \mid H) = P\left[\frac{\bar{X} - 0}{1/\sqrt{n}} > \frac{k}{1/\sqrt{n}}\right]$$

If $\alpha = 0.05$, then from the tables of the standard normal distribution, **we have**

$$\frac{k}{1/\sqrt{n}} = 1.645 \text{ or, } k = 1.645/\sqrt{n}.$$

The test at level of significance $\alpha = 0.05$ is thus: ; reject H if $\bar{X} > 1.645/\sqrt{n}$ and accept H, otherwise, i.e., if $\bar{X} \leq 1.645/\sqrt{n}$.

The probability of Type II error in this case is

$$\beta = P\left[\bar{X} \leq 1.645/\sqrt{n} \mid A\right]$$

and the power is

$$\gamma = 1 - \beta = 1 - P\left[\bar{X} \leq 1.645/\sqrt{n} \mid A\right]$$

$$= P\left[\bar{X} > 1.645/\sqrt{n} \mid A\right]$$

$$= P\left[\frac{\bar{X} - 1}{1/\sqrt{n}} > \left(\frac{1.645}{\sqrt{n}} - 1\right)\sqrt{n}\right]$$

$$= P\left[Z > 1.645 - \sqrt{n}\right]$$

Where Z is a standard normal variate. The above probability can be evaluated using tables of standard normal distribution.

How do we know in a given situation whether a test is best (in the sense of having maximum power)? Or, how do we construct a best test procedure? These and related questions will be taken up in detail in Unit 17.

E8) In the following cases, examine which of the hypotheses are simple and which are composite:-

(i) $H : p \leq p_0$ (given) where $p_0 > 0$, and X follows a Binomial distribution with known n and unknown p.

(ii) $H : p = 0.6$, where p is as in (i).

(iii) $H : \mu = \mu_0$ where μ is the mean of the normal population with unknown variance σ^2.

(iv) $H : \sigma^2 \geq 1$ where σ^2 is the variance of a normal population with mean zero.

15.7 INTERVAL ESTIMATION

In Section 15.5 of this Unit, we studied some notions about the point estimation of a parameter. In point estimation, a single value (based on the sample values) is suggested as an estimator of the parameter in question. Alternatively, one may be interested in proposing an interval or a set, of which the parameter is likely to be a member. This interval or set will depend on the observed data. Such an estimation problem is called interval estimation, or, obtaining confidence interval or confidence set for the parameter. We shall illustrate the basic theory of interval estimation with the help of a simple example.

Example 16 : Let X be a random variable having a normal distribution with mean μ (unknown) and variance unity. We know that $\overline{X}$, the sample mean of a random sample of size n from N (μ, 1) is a (point) estimator of μ. Also, the random variable $Z = (\overline{X} - \mu)/(1/\sqrt{n})$ has a standard normal distribution. Given a number α, $0 < \alpha < 1$, we can choose a and b such that

$$P[a \le Z \le b] = 1 - \alpha.$$

Since the standard normal distribution is symmetric about zero, let us choose $Z_{\alpha/2}$ such that

$$P\left[-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right] = 1 - \alpha$$

The point $Z_{\alpha/2}$ can be found from the tables of standard normal distribution. The above probability statement can be written as

$$1 - \alpha = P\left[-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\right] = P_\mu\left[-Z_{\alpha/2} \le \frac{\overline{X} - \mu}{1/\sqrt{n}} \le Z_{\alpha/2}\right]$$

where P_μ is the probability when μ is the parameter. We may again rewrite it as

$$1 - \alpha = P_\mu\left[\overline{X} - Z_{\alpha/2}/\sqrt{n} \le \mu \le \overline{X} + Z_{\alpha/2}/\sqrt{n}\right].$$

The above is a confidence statement about the unknown parameter μ. Note that while μ is a parameter and not a random variable, the end points of the interval, $(\overline{X} \pm Z_{\alpha/2}/\sqrt{n})$ are random variables, being dependent on the sample observations. The above statement is interpreted as follows: the probability that the random interval contains the parameter μ is $1 - \alpha$. The interval

$$\left[\overline{X} - Z_{\alpha/2}/\sqrt{n}, \overline{X} + Z_{\alpha/2}/\sqrt{n}\right]$$

is called a $100(1 - \alpha)$ percent confidence interval for μ and $(1 - \alpha)$ is called the confidence coefficient. The term 'confidence' is based on the fact that the statistician believes that the random interval is likely to contain the parameter μ, $100(1 - \alpha)$ times in repetitions of computing the interval from 100 random samples $\{X_i, i = 1, 2, \ldots, n\}$. We now have a formal definition.

Definition 6: Let X be a random variable with d.f. $F(x; \theta)$ here θ is a scalar parameter. Suppose $\overline{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from $F(x; \theta)$. The random interval $[r_L(\underline{X}), r_U(\underline{X})]$ is called a confidence interval for θ with

confidence coefficient $(1-\alpha)$ if

$$P_\theta [\, \theta \in r_L(\underline{X}), r_U(\underline{X}) = 1-\alpha \text{ for all } \theta \in \Omega.$$

Here $r_L(X_1, \ldots, X_n)$ and $r_U(X_1, \ldots, X_n)$, $r_L(X_1, \ldots, X_n) < r_U(X_1, \ldots, X_n)$ are two statistics such that the probability that $r_L(\underline{X})$ and $r_U(\underline{X})$ contain θ is $1-\alpha$. The interval $[\, r_L(\underline{X}), r_U(\underline{X})\,]$ is also called an interval estimator of θ.

Referring to Example 16, we have

$$r_L = \overline{X} - Z_{\alpha/2}/\sqrt{n}, r_U = \overline{X} + Z_{\alpha/2}/\sqrt{n}.$$

Therefore the length of the interval is

$$r_U - r_L = 2Z_{\alpha/2}/\sqrt{n}.$$

If we want to have confidence interval of specified length, say d, then the relation

$$d = 2 \cdot Z_{\alpha/2}/\sqrt{n}$$

should hold, or equivalently, in terms of the sample size n, we must have

$$n = (2Z_{\alpha/2}/d)^2.$$

This formula is helpful in determining the sample size n needed in order to get a confidence interval of specified length d and confidence coefficient $(1-\alpha)$. In practice, n computed from the above formula may not be an integer. If that be the case, one chooses the sample size to be the smallest integer greater than or equal to n defined by the above formula.

E9) Let X follow a normal distribution with unknown mean μ and known variance σ_0^2. It is desired to have a confidence interval for μ with confidence coefficient 0.95 and length $2\sigma_0$. What should be the sample size to achieve this?

It is obvious that $P[\theta \in (-\infty, \infty)] = 1$ for any scalar parameter θ and hence the entire real line is a confidence interval for θ with confidence coefficient 100 percent. However, this interval is too large to be of any practical use. In interval estimation, we look for confidence intervals whose lengths are as small as possible with largest confidence coefficient. We shall discuss more about this in Unit 18.

15.8 SUMMARY

In this Unit, we have

1. briefly introduced the problem of statistical inference and introduced the concept of random sampling,
2. discussed some sampling distributions of statistics based on samples from normal population,
3. introduced the problem of point estimation of parameters and discussed unbiasedness and consistency of parameters and discussed unbiasedness and consistency of estimators,
4. discussed the problem of testing of hypothesis and introduced the basic concepts like types of error, level of significance and power,
5. discussed the problem of interval estimation.

15.9 SOLUTIONS AND ANSWERS

E1) The sample values are 0, 1, 1, 1, 0. Note that since the random variable takes only two values 1 and 0 with respective probabilities p and q (= 1 – p), the sample values are 0 and 1 only. Since $n = 5, \bar{X} = n^{-1}\sum_{i=1}^{5} X_i = 3/5 = 0.6$.

Also

$$S^2 = (n-1)^{-1}\sum_{1}^{5}(X_i - \bar{X})^2 = \frac{1}{4}\sum_{i=1}^{5}(X_i - 0.6)^2$$

$$= \frac{1}{4}(0.36 + 0.16 + 0.16 + 0.16 + 0.36) = 1.20/4 = 0.3.$$

Hence, $\bar{X} = 0.6, S^2 = 0.3$.

E2) $X_1, X_2, \ldots, X_n$ is a sample from a normal population with mean zero and variance σ^2. Define $Y_i = X_i/\sigma$ for $i = 1, 2, \ldots, n$. Then, $Y_i = 1, 2, \ldots, n$ are independent (because X'_i is are so and σ is a constant) and each Y_i has a normal distribution with mean zero and variance unity. Hence

$$\sum_{i=1}^{n} X_i^2/\sigma^2 = \sum_{i=1}^{n} Y_i^2$$

follows a chi-square distribution with n D.F.

E3) Here $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance unity. $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ is the sample mean. Now, we know (cf. Section 15.4) that if $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ then $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2$ has a chi-square distribution with (n – 1) D.F. Here $\sigma^2 = 1$ and it follows that $\sum_{i=1}^{n}(X_i - \bar{X})^2$ has a chi-square distribution with (n – 1) D.F.

E4) If $X_1, X_2, \ldots, X_n$ are n independent Poisson random variables, each with the same parameter θ, then $Y = \sum_{i=1}^{n} X_i$ also has a Poisson distribution with parameter $n\theta$. Now, in the given problem, since $X_1, X_2, \ldots, X_n$ is a random sample from a Poisson distribution with parameter θ, X_i's are independent and each X_i has a Poisson distribution with parameter θ. Therefore $\sum_{i=1}^{n} X_i$ is Poisson with parameter $n\theta$. Hence

$$E(\bar{X}) = E\left(n^{-1}\sum_{i=1}^{n} X_i\right) = n^{-1}E\left(\sum_{i=1}^{n} X_i\right) = n^{1}(n\theta) = \theta,$$

and $\bar{X}$ is unbiased for θ. (Recall that the mean of a Poisson random variable with parameter θ is θ).

Also, if $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite variance σ^2, then $S^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ has expectation σ^2 (cf. Section 15.3). In the present problem, σ^2 = variance of a Poisson random variable with parameter $\theta = \theta$. Since θ is finite, S^2 is also an unbiased estimator of θ.

E5) Let $S = n\bar{X} = \sum_{i=1}^{n} X_i$. Then

$$E\left[\bar{X}(1-\bar{X})\right] = E\left[\frac{S}{n}(1-S/n)\right] = E(S/n) - E(S^2/n^2).$$

But S has a binomial distribution with parameters n and p, as it is the sum of n independent Bernoulli random variables, each with parameter p. Therefore,

$$E(S) = np, \quad Var(S) = np(1-p)$$

and $E(S^2) = Var(S) + [E(S)]^2 = np(1-p) + n^2p^2$

Hence,

$$E[\bar{X}(1-\bar{X})] = np/n - [np(1-p) + n^2p^2]/n^2$$

$$= p - p(1-p)/n - p^2$$

$$= p(1-p)(1-1/n) = p(1-p)(n-1)/n.$$

Thus, $\bar{X}(1-\bar{X})$ is not an unbiased estimator of $p(1-p)$.

E6) Since $\bar{X}$ is unbiased for μ, the first sufficient condition for consistency is automatically satisfied. Also,

$$Var(\bar{X}) = \sigma^2/n \to 0 \text{ as } n \to \infty \text{ for all finite } \sigma^2.$$

Hence the second condition is also satisfied and $\bar{X}$ is consistent for μ.

E7) Here $X_1, X_2, \ldots, X_n$ are n independent Bernoulli random variables with common probability of success, p. Hence $(X_1 + X_2 + \ldots + X_n)$ is a Binomial random variable with parameters n and p. Let $S = \sum_{i=1}^{n} X_i$ and consider the estimator $T = S/(n+1)$. Then

$$E(T) = E(S)/(n+1) = \frac{np}{n+1} = p/(1+1/n) \to p \text{ as } n \to \infty.$$

Also, $Var(T) = Var(S)/(n+1)^2 = np(1-p)/(n+1)^2 \to 0$ as $n \to \infty$.

Thus, $T = \sum_{i=1}^{n} X_i/(n+1)$ is not unbiased for p as $E(T) \neq p$, but is a consistent estimator of p.

E8) i) Composite, because H_0 contains more than one point.

ii) Simple, because H_1 has just

iii) Composite, because σ^2 is unspecified and hence $\Omega_0 = \{\mu_0, \sigma^2 > 0\}$ contains more than one point.

iv) Composite, because $\Omega_0 = \{\sigma^2 : \sigma^2 \geq 1\}$ contains more than one point.

E9) Define $Z = (\bar{X} - \mu)/(\sigma_0/\sqrt{n})$. Then, Z has a standard normal distribution. Let $Z_{\alpha/2}$ be such that

$$P\left[-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right] = 1 - \alpha, \quad 0 < \alpha < 1.$$

Here, $1 - \alpha = 0.95$ so that $\alpha = .05$ and from the tables of standard normal distribution, $Z_{\alpha/2} = 1.96$. Then length of the interval $\left[\bar{X} - Z_{\alpha/2}\,\sigma_0/\sqrt{n}, \bar{X} + Z_{\alpha/2}\,\sigma_0/\sqrt{n}\right]$ is $2Z_{\alpha/2}\,\sigma_0/\sqrt{n}$. We want the length to be $d = 2\sigma_0$. Therefore, we equate d to $2Z_{\alpha/2}\,\sigma_0/\sqrt{n}$ and solve for n. This gives $\sqrt{n} = Z_{\alpha/2}$ or $n = (Z_{\alpha/2})^2 = (1.96)^2 \sim 4$.

ADDITIONAL EXERCISES

1. Let $X_i = i$ for $i = 1, 2, \ldots, n$. Compute $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ and

   $S_0^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

2. Let X_1, X_2 be a random sample of size 2 from a df. Define

   $S_0^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. If $S_0^2 = c\,(X_1 - X_2)^2$, what is the value of c?

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Find the mean and variance of $S_0^2 = \sum_{i=1}^{n} (X_i - \bar{X}^2)/n$ where

   $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$.

4. Let X_1 and X_2 be two independent random variables with $\text{Var}(X_1) = k$, $\text{Var}(X_2) = 2$. If the variance of $Y = 3X_2 - X_1$ is 25, find k.

5. Let $X_1, X_2, \ldots, X_5$ be a random sample from the distribution with probability density function $f(x) = 6x(1 - x)$, $0 < x < 1$, zero elsewhere. If $\bar{X} = (X_1 + X_2 + \ldots + X_5)/5$, find the mean and variance of $\bar{X}$.

6. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean zero and variance θ. Show that $\sum_{i=1}^{n} X_i^2/n$ unbiased estimator of θ.

7) Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter θ, and let $\bar{X} = n^{-1} \Sigma X_i$, $S^2 = (n-1)^{-1} \Sigma (X_i - \bar{X})^2$. Show that $a\bar{X} + (1 - a)\,S^2$ with $0 \leq a \leq 1$ is unbiased for θ.

8) Let X_1 be an observation drawn at random from a distribution with probability mass function

$$f(x ; \theta) = \theta (1 - \theta)^x, \; x = 0, 1, 2, \ldots$$

$= 0$, otherwise.

Let an estimator $T(X_1)$ be defined as

$T(X_1) = 1$, if $X_1 = 0$

$= 0$, otherwise

Show that $T(X_1)$ is unbiased for θ.

9) Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution function for which the 2r-th moment about zero, μ'_{2r}, exists, $r = 1, 2, \ldots$. Let μ_r' be the r-th sample moment about zero. Show that m'_r is a consistent estimator of m_j'. μ_r'.

10) Let $X_1, X_2, \ldots, X_{25}$ be a random sample of size 25 from a normal population with mean θ and variance 100. It is desired to test the hypothesis, $H : \theta = 75$ against the alternative $A : \theta > 75$. If the test procedure is to reject H if $\overline{X} > 78$, where $\overline{X}$ is the sample mean, what is the level of significance? You are given that $P(Z \leq 2.5) = 0.994$ where Z is the standard normal variate.

11) Let $\overline{X}$ be the sample mean of a sample of 25 observations drawn from a normal population with mean θ and variance 100. Find the confidence coefficient for the confidence interval $(\overline{X} \pm 3.92)$ for θ.

12) It is desired to test whether a given coin is fair. For this purpose, we set up the hypothesis $H : p = 1/2$ against the alternative $A = : p = 3/4$ (say) where p is the probability of obtaining a head. The coin is tossed 5 times and number of heads noted. It is decided to reject the hypothesis if 5 heads show. Find the level of significance and power of the test. For facilitating the computations, the values of $f(x, p) = \binom{5}{x} p^x (1-p)^{5-x}$ are given below for $p = 1/2$ and $p = 3/4$.

x	0	1	2	3	4	5
f (x ; 1/2)	1/32	5/32	10/32	10/32	5/32	1/32
f (x ; 3/4)	1/1024	15/1024	90/1024	270/1024	405/1024	243/1024

UNIT 16 POINT ESTIMATION

Structure

16.1 INTRODUCTION

In Unit 15, you have been introduced to the problem of point estimation and also to some basic concepts of the theory of point estimation. There we have also discussed two desirable properties of an estimator, viz., unbiasedness and consistency. In this unit, the problem of point estimation will be discussed in greater detail. To begin with, we shall introduce some more concepts. Next, some methods of point estimation are discussed. In particular, we shall concentrate on two methods of estimation that are used widely in practice, viz., the method of moments and the method of maximum likelihood. The first one is easy to implement in practice and the latter leads to estimators with "good" properties.

Objectives

After reading this unit, you should be able to ;

- list the criteria for the choice of a good estimator
- derive estimators by one of the methods discussed
- decide which one in a given class of estimators is best according to a given criterion
- assess the goodness or otherwise of any given estimator.

16.2 PROPERTIES OF AN ESTIMATOR

We have already discussed in Unit 15 two properties of an estimator, namely, unbiasedness and consistency. Let us recall the definitions of unbiasedness and consistency.

Definition 1: An estimator $T(X_1, X_2, \ldots, X_n)$, which is a function of the sample values $X_1, X_2, \ldots, X_n$ is unbiased for $g(\theta)$, a known function of the parameter θ, if

$$E_\theta \left[T(X_1, X_2, \ldots, X_n) \right] = g(\theta) \text{ for all } \theta \in \Omega$$

where E_θ denotes the expectation taken when θ is the parameter and Ω is the parameter space

Definition 2: An estimator $T_n = T(X_1, X_2, \ldots, X_n)$ is said to be a consistent estimator of θ if

$$P_\theta \left[|T_n - \theta| > \varepsilon \right] \to 0 \text{ as } n \to \infty \text{ for every } \varepsilon > 0$$

In a given problem there might exist more than one unbiased estimator for the same parameter θ or the same parametric function $g(\theta)$. How do we choose among these unbiased estimators? As mentioned in Unit 15, one way to choose among various unbiased estimators for the same parameter is to compare their variances. That is, if $T_1(X_1, X_2, \ldots X_n)$ and $T_2(X_1, X_2, \ldots X_n)$ are two unbiased estimators of $g(\theta)$, then T_1 will be preferred over T_2 if

$\text{Var}_\theta(T_1) \leq \text{Var}_\theta(T_2)$ for all $\theta \in \Omega$ and with strict inequality for at least one $\theta \in \Omega$.

This brings us to the concept of **uniformly minimum variance unbiased estimators** (UMVUE). We have the following definition.

Definition 3: For a fixed sample size, n, $T = T(X_1, X_2, \ldots X_n)$ is called a minimum variance unbiased estimator of $g(\theta)$ if (i) $E_\theta(T) = g(\theta)$ for all $\theta \in \Omega$, i.e., T is unbiased for $g(\theta)$, and (ii) $\text{Var}_\theta(T) \leq \text{Var}_\theta(T')$ for all $\theta \in \Omega$ with strict inequality for at least one $\theta \in \Omega$, where T' is any other estimator based on $X_1, X_2, \ldots X_n$ satisfying (i).

How do we locate a minimum variance unbiased estimator in a given problem ? From definition 3 alone, it may be a very difficult task, if not impossible, to find a minimum variance unbiased estimator. The following example illustrates this fact.

Example 1: Suppose a random variable X follows a normal distribution with mean θ and variance unity, and let $X_1, X_2, \ldots X_{10}$ be a random sample of size 10 from the population . We know that $\bar{X}$, the sample mean, is unbiased for θ and so is X_1. Now, $\text{Var}_\theta(\bar{X}) = 1/10$, $\text{Var}_\theta(X_1) = 1$. Therefore, $\bar{X}$ is superior to X_1 for estimating θ unbiasedly. However, this does not necessarily mean that $\bar{X}$ is the minimum variance unbiased estimator of θ. To check whether $\bar{X}$ indeed is the minimum variance unbiased estimator of θ, it will be necessary to compare the variance of $\bar{X}$ with the variances of all other unbiased estimators of θ, which is clearly an impossible task. One has therefore take recourse to other methods for locating an unbiased estimator with the smallest variance in the class of all unbiased estimators.

To formalize the concepts, we now consider a population with probability density function (if the random variable in question is continuous) or probability mass function (in the discrete case) $f(x ; \theta)$ where the parameter $\theta \in \Omega \subset R$ is a scalar. The set of all x where $f(x ; \theta) \neq 0$ is called the **support** of $f(x ; \theta)$. **We shall assume that the support of** $f(x ; \theta)$ **is independent of** θ. For example, our discussion will not be applicable to a uniform distribution over the interval $(0, \theta)$, since the support $(0, \theta)$ is dependent on the parameter θ.

The problem is to estimate a parameter θ on the basis of the data $X_1, X_2, \ldots X_n$, which is a random sample of size n from $f(x ; \theta)$. At this stage, it is important to bring in the notion of a likelihood function. Let $X_1, X_2, \ldots X_n$ be a random sample from $f(x ; \theta)$ where $f(x ; \theta)$ is the probability density (or mass) function of a random variable X. The joint probability density or mass function of $X_1, X_2, \ldots X_n$ for given θ, is

$$\prod_{i=1}^{n} f(x_i ; \theta) = L_n(\theta), \text{ say,}$$

where $x_1, x_2, \ldots x_n$ are a realization of $X_1, X_2, \ldots X_n$ for the given sample. If θ is unknown and varies over Ω, $L_n(\theta)$ may be regarded as a function of the variable θ, and is called the likelihood function of θ.

We shall henceforth assume that X is continuous and hence f (x ; θ) is a probability density function. The likelihood function based on the sample $X_1, X_2, \ldots X_n$ is

$$L_n(\theta) = f(x_1 ; \theta) f(x_2 ; \theta) \ldots . f(x_n ; \theta).$$

$\underline{X} = (X_1, X_2, \ldots, X_n)$
$d\underline{x} = (dx_1, dx_2, \ldots, dx_n)$

Suppose g ($\underline{x}$) is an estimator of θ such that

$$E_\theta [\, g(\underline{X}) \,] < \infty. \text{ Let}$$

$$B(\theta) = E_\theta [\, g(\underline{X}) \,] - \theta$$

B (θ) is called the bias of the estimator g ($\underline{X}$) in estimating θ. Clearly, if g ($\underline{X}$) is unbiased for θ, then B (θ) = 0. Now,

$$\ln L_n(\theta) = \sum_{i=1}^{n} \ln f(x_i ; \theta)$$

and assuming f (x ; θ) to be differentiable w.r.t. θ.

$$\frac{d}{d\theta} \ln L_n(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta} \ln f(x_i ; \theta)$$

The function $\frac{d}{d\theta} \ln L_n(\theta)$ is called the score function based on the observations $X_1, X_2, \ldots X_n$. Now, since f (x ; θ) is a density function, we have

$$\underbrace{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty}}_{\text{n times}} f(x_1 ; \theta) f(x_2 ; \theta) \ldots . f(x_n ; \theta)\, dx_1, dx_2, \ldots, dx_n = 1$$

for all θ.

For brevity, we write the above equation as

$$\int_{R^n} \prod_{i=1}^{n} f(x_i ; \theta)\, d\underline{x} = 1 \qquad (1)$$

Since $E_\theta [\, g(\underline{X}) \,] = \theta + B(\theta)$, we have

$$\int_A g(\underline{x}) \prod_{i=1}^{n} f(x_i ; \theta)\, d\underline{x} = \theta + B(\theta) \qquad (2)$$

where A is that part of R^n where $L_n(\theta)$ is positive.

We now assume the (1) and (2) can be differentiated w.r.t. θ under the integral sign. Then.

$$\frac{d}{d\theta} \left[\int_A L_n(\theta)\, d\underline{x} \right] = \int_A \frac{d}{d\theta} L_n(\theta)\, d\underline{x} = 0 \qquad (3)$$

and

$$\frac{d}{d\theta} \left[\int_A g(\underline{x})\, L_n(\theta)\, d\underline{x} \right] = \int_A g(\underline{x}) \frac{d}{d\theta} L_n(\theta)\, d\underline{x}$$

$$= 1 + B'(\theta) \qquad (4)$$

where $B'(\theta) = \frac{d}{d\theta} B(\theta)$

Making use of the relation

$\frac{d}{d\theta} L_n(\theta) = \left(\frac{d}{d\theta} \ln L_n(\theta) \right) L_n(\theta)$, we can write (3) and (4) alternatively as

$$\int_A \left[\frac{d}{d\theta} \ln L_n(\theta) \right] L_n(\theta)\, d\underline{x} = 0 \tag{5}$$

and

$$\int_A g(\underline{x}) \frac{d}{d\theta} \left[\ln L_n(\theta) \right] L_n(\theta)\, d\underline{x} = 1 + B'(\theta) \tag{6}$$

respectively.

Since $L_n(\theta)$ is the joint density of $X_1, X_2, \ldots, X_n$ when θ is the parameter, the relations (5) and (6) may be written in terms of expectations, as

$$E_\theta \left[\frac{d}{d\theta} \ln L_n(\theta) \right] = 0 \tag{7}$$

and

$$E_\theta \left[g(\underline{X}) \frac{d}{d\theta} \ln L_n(\theta) \right] = 1 + B'(\theta) \tag{8}$$

Combining (7) and (8) we have

$$E_\theta \left[(g(\underline{X}) - \theta) \frac{d}{d\theta} \ln L_n(\theta) \right] = 1 + B'(\theta). \tag{9}$$

The Cauchy-Schwarz inequality states that for any two random variables U and V with $E(U^2) < \infty$, $E(V^2) < \infty$,

$$\left[E_\theta(UV) \right]^2 \le E_\theta(U^2) E_\theta(V^2) \tag{10}$$

with equality if and only if U and V are linearly related.

Let $U = g(\underline{X}) - \theta$, $V = \frac{d}{d\theta} \ln L_n(\theta)$

Then from (10) we have

$$\left[1 + B'(\theta) \right]^2 = \left[E_\theta \left(g(\underline{X}) - \theta \right) \frac{d}{d\theta} \ln L_n(\theta) \right]^2$$

$$\le E_\theta \left[g(\underline{X}) - \theta \right]^2 E_\theta \left[\frac{d}{d\theta} \ln L_n(\theta) \right]^2$$

or

$$E_\theta \left[g(\underline{X}) - \theta \right]^2 \ge \frac{\left[1 + B'(\theta) \right]^2}{I_n(\theta)} \tag{11}$$

where $I_n(\theta) = E_\theta \left[\frac{d}{d\theta} \ln L_n(\theta) \right]^2$. $I_n(\theta)$ is called the **Fisher information** in the sample $(X_1, X_2, \ldots, X_n)$. The equality (11) is known as the **Cramer-Rao inequality**.

It can be shown that

$$I_n(\theta) = nI_1(\theta)$$

where $I_n(\theta)$ is the Fisher information contained in one observation. The inequality (11) can then be written alternatively as

$$E[g(\underline{X}) - \theta]^2 \geq \frac{[1 + B'(\theta)]^2}{nI(\theta)} \qquad (12)$$

where, we write $I(\theta)$ in place of $I_1(\theta)$ for simplicity.

If $g(\underline{X})$ is unbiased for θ, that is, if $E_\theta(g(\underline{X})) = \theta$, then $E_\theta[g(\underline{X}) - \theta]^2 = Var_\theta[g(\underline{X})]$ and $B(\theta) = 0$ and hence $B'(\theta) = 0$. Thus, **for an unbiased estimator** $g(\underline{X})$ of θ, we have

$$Var_\theta[g(\underline{X})] \geq 1/\{nI(\theta)\} \qquad (13)$$

The lower bound $1/(nI(\theta))$ to the variance of an unbiased estimator $g(\underline{X})$ of θ, is called the **Cramer - Rao lower bound.** Thus if the regularity conditions assumed earlier hold, the variance of an unbiased estimator $g(\underline{X})$ of θ cannot be smaller than $1/(nI(\theta))$ and hence if an unbiased estimator of a θ has variance equal to $1/(nI(\theta))$. It is the minimum variance unbiased estimator of θ.

If $g(\underline{X})$ is an unbiased estimator of $\delta(\theta)$, a known function of θ, the Cramer-Rao inequality takes the form

$$Var_\theta[g(\underline{X})] \geq [\delta'(\theta)]^2/\{nI(\theta)\} \qquad (14)$$

We can now define an **efficient estimator.**

Definition 4: An unbiased estimator $g(\underline{X})$ of $\delta(\theta)$ is said to be efficient in the Cramer-Rao sense if its variance is equal to the lower bound $[\delta'(\theta)]^2/[nI(\theta)]$ where n is the sample size and $I(\theta)$ is the Fisher information in a single observation.

It is also a uniformly minimum variance unbiased estimator (UMVUE) of $\delta(\theta)$ in the sense that it has the smallest variance uniformly for all $\theta \in \Omega$ in the class of all unbiased estimators.

Note that it is possible that there exists a uniformly minimum variance unbiased estimator for $\delta(\theta)$ but the variance of this estimator does not attain the Cramer - Rao lower bound.

The Fisher information $I(\theta)$ can be shown to be equal to

$$-E_\theta\left[\frac{d^2}{d\theta^2} \ln f(x;\theta)\right].$$

This is sometimes computationally simpler compared to the formula $E_\theta\left[\frac{d}{d\theta} \ln L_n(\theta)\right]^2$, given earlier.

Example 2: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with unknown mean μ and variance unity. The density function of a normal random variable with mean μ and variance unity is

$$f(x;\mu) = (2\pi)^{-1/2} \exp\left[-\frac{1}{2}(x-\mu)^2\right]$$

and thus

$$\ln f(x;\mu) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}(x-\mu)^2,$$

$$\frac{d}{d\mu}\ln f(x;\mu) = x-\mu$$

$$\frac{d^2}{d\mu^2}\ln f(x;\mu) = -1.$$

Hence $I(\mu) = -E\left[\frac{d^2}{d\mu^2}\ln f(x;\mu)\right] = 1$ and the Cramer - Rao lower bound is $1/[\,nI(\mu)\,] = 1/n$. Now, we know that $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ is unbiased for μ and $\mathrm{Var}_\mu(\overline{X}) = 1/n$. Therefore, $\mathrm{Var}_\mu(\overline{X})$ attains the Cramer-Rao lower bound and is the UMVUE of μ. It can be shown that there is only one such UMVUE, that is, is the unique UMVUE of μ.

Example 3: Let for $n \geq 3$, $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a Poisson population with parameter λ. Then, $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ is unbiased for λ, and $\mathrm{Var}_\lambda(\overline{X}) = \lambda/n$. Now,

$$f(x;\lambda) = e^{-\lambda}\lambda^x/x!,$$

$$\ln f(x;\lambda) = -\lambda + x\ln\lambda - \ln(x!)$$

$$\frac{d}{d\lambda}\ln f(x;\lambda) = -1 + x/\lambda,$$

and

$$\frac{d^2}{d\lambda^2}\ln f(x;\lambda) = -x/\lambda^2$$

Therefore, $I(\lambda) = -E_\lambda\left[\frac{d^2}{d\lambda^2}\ln f(x;\lambda)\right] = E(x)/\lambda^2 = \lambda^{-1}$.

So that the Cramer-Rao lower bound is $1/[nI(\lambda)] = \lambda/n$. Since $\mathrm{Var}_\lambda(\overline{X}) = \lambda/n$, $\overline{X}$ is the UMVUE of λ.

E1) Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables, that is, $X_1, X_2, \ldots, X_n$ are independent random variables with $P(X_i = 1) = p$, $P(X_i = 0) = 1-p$ for $i = 1, 2, \ldots, n$. Show that if $S = X_1 + X_2 + \ldots + X_n$, S/n is the UMVUE of p.

The next example demonstrates that a uniformly minimum variance unbiased estimator for a parameter might exist but the Cramer-Rao lower bound is not attained.

Example 4: Let X be a Poisson random variable with parameter θ and suppose we wish to estimate $\delta(\theta) = e^{-\theta}$ on the basis of a sample of size one. Consider the estimator

$$T(X) = 1, \text{if } X = 0$$

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

Then, $E_\theta [T(X)] = 1 \cdot P_\theta [X = 0] = e^{-\theta}$, so that T (X) is unbiased for $e^{-\theta}$ Also,

$$\text{Var}_\theta [T(X)] = E_\theta \left[\{T((X)\}^2 \right] - \left[E_\theta \{T((X)\} \right]^2$$

$$= E_\theta [\{T(X)\}]^2 - e^{-2\theta}$$

But $E_\theta \left[\{T(X)\}^2 \right] = E_\theta [T(X)] = e^{-\theta}$ and hence

$$\text{Var}_\theta (T(X)) = e^{-\theta} - e^{-2\theta} = e^{-\theta}(1 - e^{-\theta}).$$

Now, the probability mass function of X is

$$f(x;\theta) = e^{-\theta}\theta^x/x!$$

and thus, $\ln f(x;\theta) = -\theta + x \ln \theta - \ln (x!)$,

$$\frac{d}{d\theta} \ln f(x;\theta) = -1 + x/\theta$$

$$\frac{d^2}{d\theta^2} \ln f(x;\theta) = -x/\theta^2.$$

$$\text{Hence } I(\theta) = -E_\theta \left[\frac{d}{d\theta} \ln f(x;\theta) \right] = \theta^{-2} E_\theta (x) = \theta^{-1}.$$

Also, $\delta(\theta) = e^{-\theta}$, so that $\delta'(\theta) = \frac{d}{d\theta}\delta(\theta) = -e^{-\theta}$. Hence, the Cramer-Rao lower bound to the variance of T (X), using (14), is

$$[\delta'(\theta)]^2/I(\theta) = \theta e^{2\theta}, \text{ as } n = 1.$$

But $\text{Var}_\theta [T(X)] = e^{-\theta}(1 - e^{-\theta}) > \theta e^{-2\theta}$ for $\theta > 0$. Thus, T (X), though unbiased for $\delta(\theta) = e^{-\theta}$, has a variance larger than the Cramer-Rao lower bound. However, it can be shown that T (X) is the only unbiased estimator of $\delta(\theta) = e^{-\theta}$ and hence is the UMVUE of $e^{-\theta}$.

We now bring in another important concept, namely, that of sufficient statistic and touch upon it briefly. Let X be a random variable having probability density (or, mass) function f (x ; θ) and $X_1, X_2, \ldots, X_n$ be independent observations on X that is, let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density (mass) function f (x ; θ). The joint distribution of ($X_1, X_2, \ldots, X_n$) clearly depends on θ. Is it possible to find a statistic (a function of ($X_1, X_2, \ldots, X_n$) which contains all the "information" about θ? Such a question becomes relevant when we want to summarize the available data, because storing large bodies of data is expensive and might give rise to errors of recording etc. Moreover , it is unnecessary if we are able to summarize the data without losing any "information". A statistic containing all information about θ is called a sufficient statistic. We give below a precise definition.

A statistic $T = T(X_1, X_2, \ldots, X_n)$ is said to be a sufficient statistic for the parameter θ if the conditional distribution of $(X_1, X_2, \ldots, X_n)$ given T does not depend on θ.

From the above definition, it is clear that if there is a sufficient statistic for θ, then since the conditional distribution of $X_1, X_2, \ldots, X_n$ given the sufficient statistic is independent of θ, no other function of the observations can have any additional information about θ, given the sufficient statistic.

16.3 METHODS OF ESTIMATION

In this Section, we shall discuss some common methods of finding estimators. We concentrate on two useful and commonly used methods, namely, the method of moments and the method of maximum likelihood.

16.3.1. Method of Moments

The method of moments for estimation of parameters is often used mainly because of its simplicity. The method consists in equating sample moments to population moments and solving the resulting equations to obtain the estimators.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with distribution function depending on a k-dimensional parameter $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Let

$$m_r' = n^{-1} \sum_{i=1}^{n} X_i^r \quad , \quad n = 1, 2, \ldots,$$

be the r-th sample moment. Suppose $\mu_r' = E(X^r)$ exists for $r = 1, 2, \ldots, k$. The method of moments involves solving the equation

$$m_r' = \mu_r'(\theta_1, \theta_2, \ldots, \theta_k)., \; 1 \leq r \leq k.$$

In order to estimate the k components of $\underline{\theta}$, one clearly needs to equate at least k sample moments to k population moments. However, which of the k moments are to be equated is not specified. In practice, one generally takes the first k moments. The method is now illustrated by some examples.

Example 5: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. Here, the parameter $\underline{\theta} = (\mu, \sigma^2)$ is 2-dimensional. In order to obtain the method of moments estimators of μ and σ^2, we equate the first two sample moments to the corresponding population moments, that is,

$$m_1' = n^{-1} \sum_{i=1}^{n} X_i = \bar{X} \text{ is equated to } E(X) = \mu$$

and $m_2' = n^{-1} \sum_{i=1}^{n} X_1^2$ is equated to $E(X^2) = \mu^2 + \sigma^2$.

The first of these two equations gives $\bar{X}$ as an estimator of μ; $\hat{\mu} = \bar{X}$. From the second, using $\hat{\mu} = \bar{X}$, we have an estimator of σ^2 as

$$\hat{\sigma}^2 = m_2' - \bar{X}^2$$

$$= n^{-1} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Observe that $\hat{\mu} = \bar{X}$ is an unbia... estimator of μ but $\hat{\sigma}^2$ is not unbiased for σ^2. However, both $\hat{\mu}$ and $\hat{\sigma}^2$ are co...ent estimators of μ and σ^2 respectively.

Example 6: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution with density function

$$f(x; \alpha, \beta) = \frac{1}{\beta - \alpha}, \quad \alpha \le x \le \beta$$

$$= 0, \qquad \text{elsewhere.}$$

Then, $\mu'_1 = E(X) = (\alpha + \beta)/2$, $E(X^2) = \mu'_2 = (\alpha^2 + \alpha\beta + \beta^2)/3$.

Instead of equating m'_1 to μ'_1 and m'_2 to μ'_2, we may as well equate m'_1 to μ'_1 and $m_2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ to $\mu_2 = \text{Var}(X)$. It is easy to see that

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

$$= (\beta - \alpha)^2/12.$$

Thus, the equations to be solved are

$$\bar{X} = (\alpha + \beta)/2$$

and $$n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = (\beta - \alpha)^2/12.$$

The solution of these equations give us the method of moments estimators of α and β as

$$\hat{\alpha} = \bar{X} - \left[3 \sum_{i=1}^{n} (X_i - \bar{X})^2/n \right]^{1/2}$$

$$\hat{\beta} = \bar{X} + \left[3 \sum_{i=1}^{n} (X_i - \bar{X})^2/n \right]^{1/2}$$

E2) Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Obtain two estimators of λ using the method of moments.

E3) Let $X_1, X_2, \ldots, X_N$ be a random sample of size N from a binomial population with parameters n and p, both unknown. Obtain the method of moments estimators of n and p.

As we have mentioned earlier, the method of moments is useful in practice because of its simplicity. The properties of such estimators are not established in general and have to be investigated separately for each estimator. Another method, which gives "efficient" estimators for large samples, under some reasonable conditions, is the method of maximum likelihood. We study this method in the following subsection.

16.3.2 Method of Maximum Likelihood

To appreciate this method of estimation, it is perhaps best to start with an example.

Example 7: Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a Poisson population with parameter θ. The likelihood function, based on the observations $X_1, \ldots, X_n$ is

$$L_n(\theta) = \prod_{i=1}^{n} e^{-\theta} \theta^{x_i} / x_i! \; ; \theta > 0.$$

The method of maximum likelihood consists in choosing as an estimator of θ that value of θ (say θ_0) which maximizes the likelihood function $L_n(\theta)$. θ_0 is called the maximum likelihood estimator of θ. Obviously, θ_0 depends on the observed sample $X_1, X_2, \ldots, X_n$. In order to find a maximum likelihood estimator of θ, we have to find the value of θ_0 at which $L_n(\theta)$ is maximum over the interval $(0, \infty)$, as $\theta > 0$ here. Now,

$$\ln L_n(\theta) = -n\theta + \left(\sum_{i=1}^{n} X_i \right) \ln \theta - \sum_{i=1}^{n} \ln (X_i!).$$

It is known that $\ln L_n(\theta)$ attains its maximum at a point θ_0 if and only if $L_n(\theta)$ attains its maximum at θ_0 Now,

$$\frac{d}{d\theta} \ln L_n(\theta) = -n + \sum_{i=1}^{n} X_i / \theta.$$

Therefore, $\frac{d}{d\theta} \ln L_n(\theta) \big|_{\theta = \theta_0} = 0$ provided $\theta_0 = n^{-1} \sum_{i=1}^{n} X_i$.

In order to verify whether $L_n(\theta)$ is indeed maximum at $\theta = \theta_0$, we compute the second derivative of $\ln L_n(\theta)$ at $\theta = \theta_0$ and check whether it is negative . Here,

$$\frac{d^2}{d\theta^2} \ln L_n(\theta) = -\theta^2 \sum_{i=1}^{n} X_i$$

and clearly, $\frac{d^2}{d\theta^2} \ln L_n(\theta) \big|_{\theta = \theta_0} < 0$. This shows that $L_n(\theta)$ is maximized at $\theta = \theta_0 = \sum_{i=1}^{n} X_i / n$. Since there is a unique maximum for $L_n(\theta)$ and the maximum is attained at $\theta = \theta_0 = n^{-1} \sum_{i=1}^{n} X_i$, θ_0 is the maximum likelihood estimator of θ.

We next consider an example where the parameter θ is a vector instead of a scalar as in Example 7.

Example 8: Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2, both unknown. The likelihood function is

$$L_n(\mu, \sigma^2) = (2\pi)^{-n/2} \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right] \cdot (\sigma^2)^{-n/2}$$

Thus,

$$\ln L_n(\mu, \sigma^2) = c - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

where c is a constant independent of μ and σ^2. The partial derivatives of $\ln L_n(\theta)$ w.r.t. μ and σ^2 are

$$\frac{d}{d\mu}\ln L_n(\mu, \sigma^2) = \sum_{i=1}^{n}(X_i - \mu)/\sigma^2$$

$$\frac{d}{d\sigma^2}\ln L_n(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}(X_i - \mu)^2/2\sigma^4.$$

Equating these two partial derivatives to zero, we get the likelihood equations. These equations have unique solutions

$$\hat{\mu} = \sum_{i=1}^{n} X_i/n = \bar{X},\ \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The verification of the fact that these solutions actually maximize the likelihood function is left to the reader. Hence, $\hat{\mu}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators of μ and σ^2 respectively.

E4) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density function

$$f(x;\theta) = \theta^{-1} e^{-x/\theta}, \qquad x > 0$$

$$= 0, \qquad \text{elsewhere.}$$

Find the maximum likelihood estimator of θ.

In the case of a scalar parameter, the likelihood function is a function of one variable (as in the case of Example 7) and if this function is twice differentiable in the domain of its definition, then one can use the methods of Calculus to find the maximum. However, if the parameter θ is a vector parameter, the likelihood function is a function of several variable and finding the points of maxima of such functions might be difficult in general. In such cases, special methods, depending on the problem on hand are needed. Of course, it is possible that the likelihood function may not be differentiable at all and in that case also, we might have to resort to special techniques. The following example is an illustration of such a situation.

Example 9: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution with density function

$$f(x;\theta) = 1/\theta, 0 \le X \le \theta$$

$$= 0, \quad \text{elsewhere.}$$

The likelihood function is

$$L_n(\theta) = \theta^{-n} \text{ if } 0 \le X_i \le \theta \text{ for } i = 1, 2, \ldots, n$$

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

We can write the likelihood function alternatively as

$$L_n(\theta) = \theta^{-n}, \text{ if } 0 \le x_{(n)} \le \theta$$

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

Where $x_{(n)}$ is the largest observation in the sample. The derivative of $L_n(\theta)$ does not vanish and hence, we cannot use the methods of Calculus to get a maximum likelihood estimator. However, $L_n(\theta)$ attains its maximum at $\hat{\theta} = x_{(n)}$ and $x_{(n)}$ is the unique maximum likelihood estimator of θ.

There is another way to look at the same problem. Since $L_n(\theta) = \theta^{-n}, 0 \le X_i \le \theta$ is an ever-decreasing function of θ, the maximum can be found by selecting θ as small as possible. Now, $\theta \ge X_i$ for $i = 1, 2, \ldots, n$ and in particular, $\theta \ge x_{(n)}$. Thus, $L_n(\theta)$ can be made no larger than $1/x_{(n)}^n$ and the unique maximum likelihood estimator of θ is $x_{(n)}$.

Are maximum likelihood estimators unbiased and unique in every situation?

The answer to both the above questions is in the negative. That maximum likelihood estimators need not be unbiased is demonstrated by making an appeal to Examples 8 and 9. In Example 8, we had seen that $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the maximum likelihood estimator of σ^2, the variance of a normal population with unknown mean μ. Clearly, this estimator is not unbiased for σ^2. Again, in Example 9, it was demonstrated that $x_{(n)}$, the largest observation in the sample is the maximum likelihood estimator of θ. But, it can be shown that $E_\theta\left(x_{(n)}\right) = n\theta/(n+1)$, so that $x_{(n)}$ is not unbiased for θ.

To see that maximum likelihood estimator need not be unique, consider the following example.

Example 10: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution over $\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$, where θ is unknown and $\theta \in \Omega = \{x : -\infty < x < \infty\}$. The likelihood is

$$L_n(\theta) = 1, \text{ if } \theta - 1/2 \le X_i \le \theta + 1/2 \text{ for } i = 1, 2, \ldots, n$$

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

or,

$$L_n(\theta) = 1, \text{ if } \theta - 1/2 \le \min(X_1, \ldots, X_n) \le \max(X_1, \ldots, X_n) \le \theta + 1/2$$

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

Thus, $L_n(\theta)$ attains its maximum provided

$$\theta - 1/2 \le \min(X_1, \ldots, X_n)$$

and

$$\theta + 1/2 \ge \max(X_1, \ldots, X_n),$$

or, when

$$\theta \le \min(X_1, \ldots, X_n) + 1/2$$

and $\theta \leq \max(X_1, \ldots, X_n) - 1/2.$

This means that any statistic $T(X_1, \ldots, X_n)$ satisfying

$$\max X_i - 1/2 \leq T(X_1, \ldots, X_n) \leq \min_i X_i + 1/2$$

is a maximum likelihood estimator of θ. In fact, for $0 < \alpha < 1$,

$$T(X_1, X_2, \ldots, X_n) = (\max_i X_i - 1/2) + \alpha(\min_i X_i - \max_i X_i + 1)$$

lies in the interval $\max_i X_i - 1/2 \leq T \leq \min_i X_i + 1/2$. Thus, for any α, $0 < \alpha < 1$, the above estimator is a maximum likelihood estimator of θ. In particular, for $\alpha = 1/2$, we get an estimator $T_1 = (\max_i X_i + \min_i X_i)/2$ and for $\alpha = 1/3$, we get the estimator $T_2 = (4 \max_i X_i + 2 \min_i X_i - 1)/6$.

Both T_1 and T_2 are maximum likelihood estimators of θ.

Are there any "good" properties of maximum likelihood estimators?

Before we attempt to answer this question, we introduce the concept of asymptotic efficiency. An estimator T_n based on a sample of size n for a parameter θ is said to be **asymptotically efficient** if $\lim_{n \to \infty} n \operatorname{Var}_\theta(T_n) = 1/I(\theta)$ where $I(\theta)$ is the per observation (Fisher) information. Recall that the Cramer-Rao lower bound to $\operatorname{Var}_\theta(T_n)$ is $1/\{nI(\theta)\}$, under some regularity conditions.

The important properties of maximum likelihood estimators are that under certain regularity conditions, these estimators are

(i) Consistent

(ii) Asymptotically efficient

(iii) Asymptotically normal with mean θ and variance $1/\{nI(\theta)\}$.

The third property says that for large samples, the distribution of the maximum likelihood estimator $\hat{\theta}$ of θ is approximately normal with mean θ and variance $1/\{nI(\theta)\}$.

The exact statements of the above results and their proofs are beyond the scope of this course and are therefore not given here.

15.4 SUMMARY

In this unit, we have

1. discussed some properties that an estimator should preferably possess, like unbiasedness, consistency and efficiency,
2. derived the Cramer-Rao lower bound to the variance of an estimator and demonstrated the use of this bound in finding minimum variance unbiased estimators,
3. discussed two commonly used methods of estimation, namely, the method of moments and the method of maximum likelihood.

16.5 SOLUTIONS AND ANSWERS

E1) Here, the probability mass function of the random variable, X is

$$f(x;p) = p^x (1-p)^{1-x}, X = 0, 1.$$

Therefore $\ln f(X;p) = X \ln p + (1-X) \ln (1-p)$,

$$\frac{d}{dp} \ln f(X;p) = X/p - (1-X)/(1-p),$$

and $$\frac{d^2}{dp^2} \ln f(X;p) = -X/p^2 - (1-X)/(1-p)^2.$$

$$\text{Hence } I(p) = E_p\left[-\frac{d^2}{dp^2} \ln f(X;p) \right]$$

$$= E_p\left[X/p^2 + (1-X)/(1-p)^2 \right]$$

$$= 1/p + 1/(1/p) = 1/\{p(1-p)\},$$

since $E_p(X) = p$. Therefore, the Cramer-Rao lower bound to the variance is $p(1-p)/n$. Let $S = X_1 + X_2 + \ldots + X_n$. Then S/n is unbiased for p. Also, $\text{Var}_p(S/n) = n^{-2} \sum_{i=1}^{n} \text{Var}_p(X_i) = n^{-2}[np(1-p)] = p(1-p)/n$. Hence S/n is the UMVUE of p.

E2) Here $X_1, X_2, \ldots . X_n$ is a random sample from a Poisson distribution with parameter λ. Hence $E(X_i) = \lambda$ for $i = 1, 2, \ldots, n$. Equating the sample mean to the population mean leads to the following equation:

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i = \lambda$$

which gives a moments estimator of λ as $\hat{\lambda} = \bar{X}$. Again, since $E(X^2) = \lambda^2 + \lambda$, equating the second sample moment about zero, viz., $n^{-1} \sum_{i=1}^{n} X_i^2$ to the corresponding population moment yields the equation

$$n^{-1} \sum_{i=1}^{n} X_i^2 = \lambda^2 + \lambda.$$

Since $\lambda > 0$, a unique positive solution of the above equation gives the second moments estimator of λ as

$$\hat{\lambda} = \left[-1 + \left[(4/n) \sum_{i=1}^{n} X_i^2 + 1 \right]^{1/2} \right]/2$$

E3) We are given that $X_1, X_2, \ldots . X_N$ is a random sample from a binomial population with parameters n and p, both unknown. We know that if X has a binomial distribution with parameters n and p, then

$$E_p(X) = np, \text{Var}_p(X) = np(1-p).$$

Therefore, $E_p(X^2) = \text{Var}_p(X) + \left(E_p(X)\right)^2 = np(1-p) + n^2 p^2$. If we

equate the first two sample moments $N^{-1}\sum_{i=1}^{N} X_i$ and $N^{-1}\sum_{i=1}^{N} X_i^2$ to the first two population moments, the following equations result:

$$\bar{X} = N^{-1}\sum_{i=1}^{N} X_i = np$$

$$s_0^2 = N^{-1}\sum_{i=1}^{N} X_i^2 = np(1-p) + n^2p^2.$$

The first of these gives $\hat{p} = \bar{X}/n$ as an estimator of p, where $\hat{n}$ is an estimator of n. Using this estimator in the second equation and solving for n gives

$$\hat{n} = \frac{\bar{X}^2}{\bar{X}^2 + \bar{X} - s_0^2} = \frac{\bar{X}^2}{\bar{X}^2 + \bar{X} - N^{-1}\sum_{i=1}^{n} X_i^2}$$

E4) Here $X_1, X_2, \ldots . X_n$ is a random sample from a population with density function

$$f(X;\theta) = \theta^{-1}\exp(-X/\theta), X > 0, \theta > 0$$

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

Therefore, the likelihood function is

$$L_n(\theta) = \theta^{-n}\exp\left(-\sum_{i=1}^{n} X_i/\theta\right)$$

$$\ln L_n(\theta) = -n\ln\theta - \sum_{x=1}^{n} X_i/\theta$$

and $\frac{d}{d\theta}\ln L_n(\theta) = -n/\theta + \sum_{x=1}^{n} X_i/\theta^2.$

Equating $\frac{d}{d\theta}\ln L_n(\theta)$ to zero, gives on solving for θ,

$\hat{\theta} = \sum_{i=1}^{n} X_i/n = \bar{X}$ the sample mean.

Also,

$$\frac{d^2}{d\theta^2}\ln L_n(\theta) = -n/\theta^2 - 2\sum_{i=1}^{n} X_i/\theta^3$$

which is negative at $\theta = \hat{\theta} = \bar{X}$. Hence $\bar{X}$ is the maximum likelihood estimator of θ.

16.6 ADDITIONAL EXERCISES

1. Let $X_1, X_2, \ldots . X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \theta^{-1} e^{-x/\theta}, \theta > 0, \text{ if } x > 0$$
$$= 0, \text{ otherwise.}$$

Show that $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$ is unbiased for θ and $\text{Var}_\theta(\overline{X}) = \theta^2/n$.

Does $\text{Var}_\theta(\overline{X})$ attain the Cramer-Rao lower bound?

2. Let $X_1, X_2, \ldots . X_n$ be a random sample from a normal population with mean zero and variance σ^2. Construct an unbiased estimator of σ as a function of $\sum_{i=1}^{n} |X_i|$. You are given that if X is normal with mean zero and variance σ^2,

$$E(|X|) = \sigma\sqrt{\frac{2}{\pi}}.$$

3. Let $X_1, X_2, \ldots . X_n$ be a random sample from a distribution having finite mean μ and finite variance σ^2. Show that $T(X_1, X_2, \ldots . X_n) = \frac{2}{n(n+1)} \sum_{i=1}^{n} i\, X_i$ is unbiased for μ.

4. Let $X_1, X_2, \ldots . X_n$ be a random sample of size n from a distribution with probability density function

$$f(X;\theta) = \theta X^{\theta-1}, 0 < X < 1, \theta > 0$$
$$= 0, \text{ elsewhere.}$$

Obtain a maximum likelihood estimator of θ.

UNIT 17 TESTING OF HYPOTHESES

Structure

17.1 INTRODUCTION

In Unit 15, we introduced some basic notions about testing of hypothesis. There we described some concepts and definitions useful in testing of hypothesis problems. In this unit, we shall discuss the problem of testing of hypothesis in greater detail. To begin with, we shall introduce some concepts and definitions. Next, we shall describe, an important result (Neyman-Pearson Lemma) for constructing critical regions for testing a simple hypothesis against a simple alternative. We also discuss the likelihood ratio test. The usage of these two procedures of testing are illustrated.

Objectives

After reading this unit, you should be able to:

- derive critical regions for testing of hypothesis,
- derive the power of these tests.

17.2 SOME CONCEPTS

In Unit 15, Section 15.5, you have been introduced to some basic notions about testing of hypothesis, like two types of error, level of significance, power critical region etc. We recall these concepts.

Let $X_1, \ldots, X_n$ be a random sample with joint distribution function $F(\underline{x}, \underline{\theta}), \underline{\theta} \in \Omega$. On the basis of the observed sample we wish to test the null hypothesis $H_0 : \underline{\theta} \in \Omega_0$ against an alternative $H_1 : \underline{\theta} \in \Omega_1 = \Omega - \Omega_0$. Both H_0 and H_1 may be simple or composite hypotheses. Let X^n, the set of all possible values of $X_1, \ldots, X_n$, denote the sample space. Then $X^n \subseteq R^n$. A rule that specifies a subset C, $C \subset X^n$, such that

if $(X_1, \ldots, X_n) \in C$, reject H_0

if $(X_1, \ldots, X_n) \notin C$ do not reject H_0

(or, accept H_0)

is called a test of H_0 against H_1 and C is called a critical region of the test. The statistic used in the specification of C is called a test statistic. In such a test procedure, one might commit two types of error. The probability of type I error is

$$\alpha(\underline{\theta}) = P_\theta(\text{Reject } H_0), \text{ when } \underline{\theta} \in \Omega_0$$

and probability of type II error is

$$\beta(\underline{\theta}) = P_\theta(\text{accept } H_0), \text{ when } \underline{\theta} \in \Omega_1. \text{ Let}$$

$$\gamma(\underline{\theta}) = P_\theta(\text{reject } H_0), \text{ when } \underline{\theta} \in \Omega_1$$

$$= P_\theta(C) \text{ when } \underline{\theta} \in \Omega_1$$

$$= 1 - \beta(\underline{\theta}) \text{ when } \underline{\theta} \in \Omega_1.$$

The function $\gamma(\underline{\theta}) = 1 - \beta(\underline{\theta})$ as a function of θ is called the power function of the test. In the construction of a test procedure, we fix the probability of the Type I error to a desired small level and choose one for which $\gamma(\underline{\theta})$ is maximum (or, equivalently, $\beta(\underline{\theta})$ is minimum). Thus, given $0 \le \alpha \le 1$, our interest is then to construct a test procedure for which

$$\alpha(\underline{\theta}) \le \alpha \text{ for } \underline{\theta} \in \Omega_0$$

and $\gamma(\underline{\theta}) = 1 - \beta(\underline{\theta})$ is as large as possible (or $\beta(\underline{\theta})$) is as small as possible, $\underline{\theta} \in \Omega_1$). A test of null hypothesis $H_0 = \underline{\theta} \in \Omega_0$ against $H_1 = \underline{\theta} \in \Omega_1$ is said to have size α, $0 \le \alpha \le 1$, if

$$\sup_{\theta \in \Omega_0} \alpha(\theta) = \alpha.$$

The chosen size α is generally unattainable. In fact in many problems only countable number of levels α in $[0, 1]$ are attainable. In such a case we usually take the largest level less than α that is attainable. We also call α as the level of significance of the critical region C if

$$\alpha(\theta) \le \alpha \text{ for all } \theta \in \Omega_0.$$

If $\sup_{\theta \in \Omega_0} \alpha(\theta) = \alpha$, then the level of significance and size of critical region, C, both equal α. On the other hand, if $\sup_{\theta \in \Omega_0} \alpha(\theta) < \alpha$ then the size of critical region C is smaller than its level of significance α. If H_0 is a simple hypothesis, then it is clear that $\alpha(\theta)$, $\theta \in \Omega_0$ is the size of the critical region C, which may or may not equal a given significance level α.

We now define a criterion for selecting a test statistic for testing $H_0 : \theta \in \Omega_0$ against $H_1 : \theta \in \Omega_1$, if H_1 is a composite hypothesis. A test with critical region C_0 of size $\alpha = \sup_{\theta \in \Omega_0} \alpha(\theta)$ is said to be **Uniformly Most Powerful** of size α of testing H_0 if it has the maximum power among all critical regions C of the same size. In other words, C_0 is the best (Uniformly Most Powerful) if for all tests C with size α (which is the size of C_0), the inequality $P_\theta(C_0) \ge P_\theta(C)$

holds for each $\theta \in \Omega_1$

Uniformly most powerful tests do not exist for many hypothesis testing problems. Even when they do exist, they are often not easy to find. We now describe a test procedure (equivalently, obtain a critical region) for testing a simple hypothesis

against a simple alternative. In this case, the power function, $\gamma(\theta)$ reduces to a single number, so that the "uniformly", in uniformly most powerful, becomes redundant, and we examine the question of the existence of a **most powerful test** of given significance level α.

17.3 NEYMAN-PEARSON LEMMA

We first state (without proof) an important result, called Neyman-Pearson Lemma which is very useful for constructing uniformly most powerful tests.

Lemma 1: Let $f_0, f_1, \ldots$ be integrable functions of $X_1, \ldots, X_n$ over a space S and let C be any region such that

$$\int_C f_i \, dx = a_i \text{ (given)}, \; i = 1, 2, \ldots \qquad \ldots (1)$$

Further, let there exist constants $k_1, k_2, \ldots$ such that for the region C_0 within which $f_0 \geq k_1 f_1 + k_2 f_2 + \ldots\ldots$ outside which $f_0 < k_1 f_1 + k_2 f_2 + \ldots\ldots$,

and the conditions (1) are satisfied. Then

$$\int_{C_0} f_0 \, dx \geq \int_C f_0 \, dx \qquad \ldots (2)$$

We now describe the application of Lemma 1 to the problem of testing of **simple** hypothesis against a **simple** alternative.

For a fixed positive integer n, let $X_1, \ldots\ldots X_n$ denote a random sample from a density $f(\underline{x}, \theta)$. Let $X_1, \ldots\ldots X_n$ denote the observed sample. Then the density of $\underline{X} = (X_1, \ldots\ldots X_n)$ is $P_\theta = \prod_{j=1}^{n} f(X_j, \theta)$. Let $P_{\theta_0}(\underline{x})$ and $P_{\theta_1}(\underline{x})$ be the densities of $\underline{x}$ under H_0 and H_1 respectively. The problem is that of determining a critical region C_0 such that

$$\int_{C_0} P_{\theta_0}(\underline{x}) \, d\underline{x} = \alpha \text{ (assigned value)} \qquad \ldots (3)$$

and

$$\int_{C_0} P_{\theta_1}(\underline{x}) \, d\underline{x} \text{ is a maximum.} \qquad \ldots (4)$$

The optimum region C_0 is provided by choosing $f_0 = P_{\theta_1}(\underline{x})$ and $f_1 = P_{\theta_0}(\underline{x})$ in Lemma 1. The optimum region C_0 is defined by

$$C_0 = \left\{ \underline{x} \,|\, P_{\theta_1}(\underline{x}) \geq k \, P_{\theta_0}(\underline{x}) \right\} \qquad \ldots (5)$$

provided there exists a **k such that (5) is satisfied.** The test can thus be written as

$$T = \frac{P_{\theta_1}(\underline{x})}{P_{\theta_0}(\underline{x})} \geq k. \qquad \ldots (6)$$

Thus we determine the distribution of T under H_0. If the distribution is continuous then there exists a k such that

$$P_{\theta_0}(T \geq k) = \alpha$$

for any assigned α. The test, $T \geq k$, depends on the simple alternative H_1. If the test is independent of the alternative hypothesis in a class of alternatives, then we have a uniformly most powerful test with respect to all such alternative hypothesis against a simple hypothesis.

We now consider some examples.

Example 1 : Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. Assume that σ^2 is fixed and known and $\mu \in (-\infty, \infty)$. We wish to obtain a critical region for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, where μ_0 and μ_1 are the specified values of μ.

We have,

$$P_{\mu_1}(\underline{X}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_1)^2\right\}$$

and

$$P_{\mu_0}(\underline{X}) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right\}.$$

The critical region, using Neyman-Pearson Lemma, is obtained as follows :

$$P_{\mu_1}(\underline{x}) \geq k\, P_{\mu_0}(\underline{x})$$

$$\Rightarrow \quad \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_1)^2\right\}$$

$$\geq \quad k \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right\}$$

$$\Rightarrow \quad \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\mu_1)^2 - \sum_{i=1}^{n}(X_i-\mu_0)^2\right]\right\} \geq k$$

$$\Rightarrow \bar{X}(\mu_1-\mu_0) \geq k_0 \text{ (say)}$$

taking natural logarithm and simplifying where $\bar{X} = 1/n \sum_{i}^{n} X_i$.

Case I

Let $\mu_1 > \mu_0$. Then the critical region is

$$\bar{X} \geq k_1 \text{ (say)}$$

where k_1 is to be determined such that

$$P_{\mu_0}\left\{\bar{X} \geq k_1\right\} = \alpha$$

or $$P\left\{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \geq \frac{k_1-\mu_0}{\sigma/\sqrt{n}}\right\} = \alpha$$

or $$P\left\{Z \geq \frac{k_1-\mu_0}{\sigma/\sqrt{n}}\right\} = \alpha$$

where Z is distributed as N (0, 1). Therefore choose k_1 so that

$$P(Z \geq Z_\alpha \mid \mu_0) = \int_{\frac{(k_1-\mu_0)}{\sigma/\sqrt{n}}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] dx = \alpha$$

Denoting by Z_α the upper α probability point of N (0, 1) distribution, we have

$$(k_1-\mu_0)/\sigma/\sqrt{n} = Z_\alpha$$

$$\Rightarrow \quad k_1 = \mu_0 + \sigma/\sqrt{n}\, Z_\alpha.$$

Hence the best critical region in this case is

$$C_0 = \left\{X \mid \bar{X} > \mu_0 + \sigma/\sqrt{n}\, Z_\alpha\right\}$$

$$\text{or } C_0 = \left\{X \mid \sqrt{n}\,(\bar{X}-\mu_0)/\sigma > Z_\alpha\right\}$$

Case II Let $\mu_1 < \mu_0$. Then

$$\bar{X} \leq k_2$$

where k_2 is choosen such that

$$P\left[\bar{X} \leq k_2 \mid \mu_0\right] = \alpha.$$

The best critical region in this case is

$$C_0 = \left\{X \mid \bar{X} < \mu_0 - Z_\alpha\ \sigma/\sqrt{n}\right\}$$

or $$C_0 = \left\{X \mid \sqrt{n}\,\frac{(\bar{X}-\mu_0)}{\sigma} < -Z_\alpha\right\}.$$

where $-Z_\alpha = Z_{1-\alpha}$ is the lower α probability point of the standard normal distribution.

It may be seen that the test $\bar{X} \geq k_1$ ($\bar{X} \leq k_2$) is uniformly most powerful for the class of alternatives $\mu_1 > \mu_0$ ($\mu_1 < \mu_0$) because k_1 (k_2) is independent of μ. But there is no uniformly most powerful test for the entire class of alternatives $\mu_1 \neq \mu_0$.

In case I, the power of the test is

$$P_{\mu_1}(C_0) = P_{\mu_1}(\bar{X} \geq k_1)$$

$$= P\left[\frac{\bar{X}-\mu_1}{\sigma/\sqrt{n}} \geq \frac{k_1-\mu_1}{\sigma/\sqrt{n}}\right]$$

$$= P\left[Z \geq \frac{k_1 - \mu_1}{\sigma/\sqrt{n}}\right]$$

$$= 1 - P\left[Z \geq \frac{k_1 - \mu_1}{\sigma/\sqrt{n}}\right]$$

$$= 1 - \phi\left[\frac{k_1 - \mu_1}{\sigma/\sqrt{n}}\right]$$

where $\phi(\cdot)$ is the distribution function of a standard normal distribution.

Similarly the power of the test for the second case is

$$P_{\mu_1}(C_0) = P_{\mu_1}\left[\bar{X} \leq k_2\right]$$

$$= P\left[Z \leq \frac{k_2 - \mu_1}{\sigma/\sqrt{n}}\right]$$

$$= \phi\left[\frac{k_2 - \mu_1}{\sigma/\sqrt{n}}\right]$$

E1) Let X_1 be a random sample of size 1 from a population with p.d.f. $f(x, \theta) = (1/\theta)\exp\left(-\frac{x}{\theta}\right)$, $x \geq 0, \theta > 0$. Obtain a best critical region of size α for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 \neq \theta_0$ and also the power of the test.

E2) Obtain a test, the size of the test and power of the test for testing a null hypothesis $H_0 : X \sim \frac{1}{\sqrt{2\pi}}\exp\left(-x^2/2\right), x \in \mathbf{R}$ against an alternative $H_1 : X \sim 1/2\ \exp\{-|x|\}, x \in \mathbf{R}$. Develop the test on the basis of a single observation.

17.4 LIKELIHOOD RATIO TESTS

In Section 17.3 we described the Neyman-Pearson lemma for obtaining the best test for testing a simple hypothesis against a simple hypothesis. But when the hypothesis to be tested is composite rather than simple, it becomes necessary to introduce some other principle for obtaining good tests.

Neyman and Pearson suggested a simple method of construction of a test statistic which is closely related to the maximum likelihood method of estimation.

Suppose $L(\theta \mid \underline{X})$ is the likelihood function of $\underline{\theta}$ corresponding to the set of values $\underline{X} = (X_1, X_2, \ldots, X_n)$. Suppose we are required to test the simple hypothesis

$$H_0 : \theta = \theta_0$$

against the composite hypothesis

$$H_1 : \theta \neq \theta_0.$$

In this situation, given the observation X, intuitively we should reject H_0 in case $L(\theta_0 | \underline{X})$ is too small and accept it otherwise. This means that the test be based on the critical region

$$C_0 = \{\underline{X} | \lambda(X) < \lambda_0\}$$

where λ is the likelihood ratio defined by

$$\lambda(\underline{X}) = \frac{L(\theta_0 | \underline{X})}{\sup_{\theta \in \Omega} L(\theta | \underline{X})}$$

and λ_0 is a constant so chosen as to make the probability of Type I error associated with the test equal to α.

When H_0 itself is a composite hypothesis, say

$$H_0 : \theta \in \Omega_0$$

the likelihood is not a constant under H_0, and in order to judge the acceptability of the null hypothesis in the light of the observation X, we compare the highest value of the likelihood under H_0, i.e.

$$\sup_{\theta \in \Omega_0} L(\underline{\theta} | \underline{X})$$

with its highest value under the model, i.e.,

$$\sup_{\theta \in \Omega} L(\underline{\theta} | X)$$

Thus here we base our test on the likelihood ratio

$$\lambda(\underline{X}) = \frac{\sup_{\theta \in \Omega_0} L(\underline{\theta} | \underline{X})}{\sup_{\theta \in \Omega} L(\underline{\theta} | \underline{X})}$$

The critical region of the size-α likelihood ration test of H_0 against H_1 is

$$C_0 = \{\underline{X} | \lambda(\underline{X}) < \lambda_0\}$$

where λ_0 is determined by the condition

$$\sup_{\theta \in \Omega_0} P_\theta \{\underline{X} | \lambda(\underline{X}) < \lambda_0\} = \alpha$$

The critical value of the ratio is determined by consideration of the size of the test.

It is clear that $0 \leq \lambda \leq 1$. As in the case of Neyman-Pearson lemma, if the distribution of λ is continuous as, then any size α is attainable. If, however, the distribution of λ is discrete it is difficult to find a likelihood ratio test whose size is exactly equal to α. It is, however, possible to obtain a likelihood ratio test of size α by **using a randomization procedure** which we shall not discuss here. We may also choose the largest C such that

$$P_\theta \{X | \lambda(X) < c\} \leq \alpha \text{ for all } \theta \in \Omega_0.$$

Example 2: We consider here the problem of testing $H_0 : \mu = \mu_0$ against all its alternatives in sampling from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. In this case

$$\Omega_0 = \left\{ (\mu_0, \sigma^2) ; \sigma^2 > 0 \right\} \text{ and}$$

$$\Omega = \left\{ (\mu_0, \mu, \sigma^2) ; -\infty < \mu < \infty, \sigma^2 > 0 \right\}.$$

We shall write $\underline{\theta} = (\mu, \sigma^2)$.

$$\sup_{\underline{\theta} \in \Omega_0} L(\underline{\theta} | \underline{X}) = \sup_{\underline{\theta} \in \Omega_0} f(X_1, X_2, \ldots, X_n, \underline{\theta})$$

$$= \sup \left[\frac{1}{\sigma (\sqrt{2\pi})^n} \exp \left\{ \sum_i^n \frac{(X_i - \mu_0)^2}{2\sigma^2} \right\} \right].$$

Under H_0, the MLE of σ^2 is $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_0)^2$.

$$\text{Thus, } \sup_{\theta \in \Omega_0} L(\underline{\theta} | \underline{X}) = \frac{e^{-n/2}}{(2\pi/n)^{n/2} \left\{ \sum_{i=1}^{n} (X_i - \mu_0^2) \right\}^{n/2}}$$

Now

$$\sup_{\theta \in \Omega} L(\underline{\theta} | \underline{X}) = \sup_{\theta \in \Omega} f(X_1, X_2, \ldots, X_n, \underline{\theta})$$

$$= \sup \left[\frac{1}{(\sigma \sqrt{2\pi})^n} \exp \left\{ \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2} \right\} \right].$$

Under H, the MLE of μ, σ^2 are

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \; ; \; \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Thus

$$\sup_{\theta \in \Omega_0} L(\underline{\theta} | \underline{X}) = \frac{e^{-n/2}}{(2\pi/n)^{n/2} \left\{ \sum_{1}^{n} (X_i - \bar{X})^2 \right\}^{n/2}}$$

$$\lambda(x) = \left[\frac{\sum_{1}^{n} (X_i - \bar{X})^2}{\sum_{1}^{n} (X_i - \mu_0)^2} \right]^{n/2}$$

$$= \left[\frac{\sum_1^n (X_i - \bar{x})^2}{\sum_1^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2} \right]^{n/2}$$

$$= \left[\frac{1}{1 + \left[n(\bar{X} - \mu_0)^2 / \sum_1^n (X_i - \bar{X})^2 \right]} \right]^{n/2}$$

The likelihood ratio rejects H_0 if

$$\lambda(\bar{X}) < c$$

$$\Rightarrow \quad \frac{1}{1 + \dfrac{n(\bar{X} - \mu_0)^2}{\sum_1^n (X_i - \bar{X})^2}} < c^{2/n}$$

Since $\lambda(X)$ is a decreasing function of $n(\bar{X} - \mu_0)^2 / \sum_1^n (X_i - \bar{X})^2$, we reject H_0 if

$$\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{\Sigma (X_i - \bar{X})^2}} \right| > C_1$$

that is, if

$$\left| \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right| > C_2,$$

where $S^2 = (n-1)^{-1} \sum_1^n (X_i - \bar{X})^2$. The

statistic $t(\underline{X}) = \dfrac{\sqrt{n}(\bar{X} - \mu_0)}{S}$

has a Student's t distribution with $(n-1)$ d.f under $H_0 : \mu = \mu_0$, but under $H_1 : \mu \neq \mu_0$, $t(\underline{X})$ **has a non-central t-distribution** with $(n-1)$ d.f. and non-centrality parameter $\delta = (\mu - \mu_0)/\sigma$. Thus the critical region is

$|t| > C_2$, (for simplicity, we write t for $t(\underline{X})$)

where C_2 is so chosen that

$$P_\theta \{ |t| > C_2 \} = \alpha.$$

Let $C_2 = t_{n-1, \alpha/2}$ in accordance with the distribution of $t(\underline{X})$ under H_0. Thus the two sided test obtained here is

$$|t| > t_{n-1,\alpha/2}.$$

Suppose we now consider the problem of testing $H_0 : \mu = \mu_0$ against the class of alternatives $H_1 : \mu = \mu_1 > \mu_0$. In this case

$$\Omega_0 = \left\{(\mu_0, \sigma^2) \mid \sigma^2 > 0\right\} \text{ and}$$

$$\Omega = \left\{(\mu_1, \sigma^2); \mu_1 > \mu_0, \sigma^2 > 0\right\}$$

$$\sup_{\theta \in \Omega_0} L(\theta \mid \underline{X}) = \frac{e^{-n/2}}{(2\pi/n)^{n/2}\left\{\sum_1^n (X_i - \mu_0)^2\right\}^{n/2}}$$

and

$$\sup_{\theta \in \Omega_0} L(\theta \mid \underline{X}) = \sup_{\substack{\mu > \mu_0 \\ \sigma^2 > 0}} \left[\frac{1}{\sigma(\sqrt{2\pi})^n} \exp - \left[\frac{-\sum_i^n (X_i - \mu)^2}{2\sigma^2}\right]\right].$$

The MLE of μ is $\overline{X}$, when $\overline{X} \geq \mu_0$ and is μ_0 if $\overline{X} < \mu_0$. Similarly, the MLE of σ^2 is $\frac{1}{n}\sum_1^n (X_i - \overline{X})^2$ when $\overline{X} \geq \mu_0$ and is $\frac{1}{n}\sum_1^n (X_i - \mu_0)^2$ when $\overline{X} < \mu_0$.

Thus

$$\sup_{\theta \in \Omega} L(\theta \mid \underline{X}) = \frac{1}{(2\pi/n)^{n/2}\left\{\sum_1^n (X_i - \overline{X})^2\right\}^{n/2}} e^{-n/2} \text{ if } \overline{X} \geq \mu_0$$

$$= \frac{1}{(2\pi/n)^{n/2}\left\{\sum_1^n (X_i - \mu_0)^2\right\}^{n/2}} e^{-n/2} \text{ if } \overline{X} < \mu_0$$

Thus

$$\lambda(\underline{X}) = \left[\frac{\sum_1^n (X_i - \overline{X})^2}{\sum_1^n (X_i - \mu_0)^2}\right]^{n/2}, \text{ if } \overline{X} \geq \mu_0$$

$$= 1 \quad \text{if } \overline{X} < \mu_0$$

Thus the observation $\underline{X}$, for which $\overline{X} < \mu_0$ fall under acceptable region.

Hence we consider those x for which $X \geq \mu_0$. Proceeding on the same lines as for the set of alternatives $H_1 : \mu \neq \mu_0$, we get the test as

$$t = \frac{\sqrt{n}\,(\bar{X} - \mu_0)}{S} > t_{n-1,\,\alpha}$$

to reject the null hypothesis. The one sided test is UMP.

In the preceeding illustrations, $\lambda(X)$ was a simple function of $\bar{X}$ and S^2 whose distribution is known. In general, however, there is no guarantee that some such nice relationship to a familiar variable will exist. Then we must use whatever tools available to find the distribution of λ. Fortunately, for large samples there is a good approximation to the distribution of λ which eliminates the necessity for finding the distribution of λ in situations where this is difficult to find. Under certain regularity conditions, the random variable $-2 \log_e \lambda$ has an asymptotic χ^2-distribution. The degrees of freedom equals the number of unknown parameters under Ω minus the number of unknown parameters under Ω_0.

E3) Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli distribution with parameter p, $0 \leq p \leq 1$. Construct a level α likelihood ratio test of $H_0 : p \leq p_0$ against $H_1 : p > p_0$.

17.5 SUMMARY

In this unit we have

1. briefly introduced the problem of testing of hypothesis,
2. discussed the Neyman-Pearson Lemma for testing a simple hypothesis against a simple alternative,
3. described the likelihood ratio test for testing hypothesis

17.6 SOLUTIONS AND ANSWERS

E1) We have

$$P_{\theta_1}(X) = \frac{1}{\theta_1} \exp\left[-X_1/\theta_1\right]$$

and

$$P_{\theta_0}(X) = \frac{1}{\theta_0} \exp\left[-X_1/\theta_0\right].$$

Using Neyman-Pearson Lemma, the best critical region is obtained as

$$\frac{P_{\theta_1}(X)}{P_{\theta_0}(X)} \geq k$$

$$\Rightarrow \quad \theta_0/\theta_1 \exp\left\{ x_1 \left(1/\theta_1 - 1/\theta_0 \right) \right\} \geq k$$

After taking logarithms, we have

$$X_1 (\theta_1 - \theta_0) \geq k_1$$

CASE I: Let $\theta_1 > \theta_0$

The test is

$$X_1 \geq k_2$$

where k_2 is to be determined such that

$$P_{\theta_0}\left[X_1 \geq k_2\right] = \alpha$$

that is, $\frac{1}{\theta_0}\int_{k_2}^{\infty} \exp(-x_1/\theta_0)\, dx_1 = \alpha$

or $\exp(-k_2/\theta_0) = \alpha$

$\Rightarrow k_2 = \log(1/\alpha)^{\theta_0}$

The critical region is thus

$$C_0 = \left\{X \mid X_1 \geq \log(1/\alpha)^{\theta_0}\right\}$$

The power of the test is

$$P_{\theta_1}(C_\theta) = P_{\theta_1}\left[X_1 \geq \log(1/\alpha)^{\theta_0}\right]$$

$$= \frac{1}{\theta_1}\int_{\log(1/\alpha)^{\theta_0}}^{\infty} \exp(-X_1/\theta_1)\, dx_1$$

$$= \left[-\exp(-X_1/\theta_1)\right]_{\log(1/\alpha)^{\theta_0}}^{\infty}$$

$$= \exp\left[-\log(1/\alpha)^{\theta_0/\theta_1}\right]$$

$$= \alpha^{\theta_0/\theta_1}$$

Case II

Let $\theta_1 < \theta_0$

The test is

$$X_1 \leq k_3$$

where k_3 is to determined such that

$$P_{\theta_0}(X_1 \leq k_3) = \alpha$$

that is

$$\frac{1}{\theta_0}\int_{0}^{k_3} \exp(-X_1/\theta_0)\, dx_1 = \alpha$$

$$\Rightarrow \left[-\exp(-X_1/\theta_0)\right]_0^{k_3} = \alpha$$

$$\Rightarrow 1 - \exp(-k_3/\theta_0) = \alpha$$

$$\Rightarrow k_3 = \log(1-\alpha)^{-\theta_0}$$

The critical region is thus

$$C_0 = \{X_1 \mid X_1 \leq \log(1-\alpha)^{-\theta_0}\}$$

The power of the test is

$$P_{\theta_1}(C_0) = \frac{1}{\theta_1}\int_0^{k_3} \exp(-X_1/\theta_1)\,dx_1$$

$$= 1 - \exp(-k_3/\theta_1)$$

$$= 1 - \exp\left\{-\frac{1}{\theta_1} - \log(1-\alpha)^{-\theta_0}\right\}$$

$$= 1 - \exp\left\{\log(1-\alpha)^{\theta_0/\theta_1}\right\}$$

$$= 1 - (1-\alpha)^{\theta_0/\theta_1}$$

E2) Since both the densities (under H_0 and H_1) are completely specified, it is a case of testing a simple hypothesis against a simple alternative. Using Neyman-Pearson Lemma, the test is obtained as

$$\frac{P_{\theta_1}(X)}{P_{\theta_0}(X)} \geq k$$

Let

$$T(X) = \frac{P_{\theta_1}(X)}{P_{\theta_0}(X)} = \frac{\sqrt{2\pi}}{2}\exp\left\{-|X| + \frac{X^2}{2}\right\}$$

$$= \sqrt{(\pi/2)}\exp\left\{\frac{1}{2}(X^2 - 2|X| + 1 - 1)\right\}$$

$$= \sqrt{(\pi/2)}\exp\left\{\frac{1}{2}\left[(|X|-1)^2 - 1\right]\right\}$$

$$= \sqrt{(\pi/2)}\exp\left(-\frac{1}{2}\right)\exp\left\{(|X|-1)^2/2\right\}$$

It is clearly seen that T (x) is an increasing function of $\big||X|-1\big|$. Hence $T(x) \geq k$ if and only if $\big||X|-1\big| \geq k'$. It **therefore follows that C_0 is of the form**

$$C_0 = \left\{X \;\middle|\; |X| \geq k_1 \text{ or } |X| \leq k_2\right\},$$

which means that if either a very large or a very small value of X is observed, then we suspect that H_1 is true and H_0 is false. The size of the test is

$$P_{\theta_0}[T(X) \geq k] = \int_{|X| \geq k_1} \frac{1}{\sqrt{2\pi}}\exp(-X^2/2)\,dx + \int_{|X| \leq k_2} \frac{1}{\sqrt{2\pi}}\exp(-X^2/2)\,dx$$

$$= 2\left\{P(Z \geq k_1) + P(0 < Z \leq k_2)\right\}, \text{ where Z}$$

is the standard normal variate. The power of the test is

$$P_{\theta_1}[T(X) \geq k] = \int_{|x| \geq k_1} \frac{1}{2} \exp(-|X|)\, dx + \int_{|x| \leq k_2} \frac{1}{2} \exp(-|X|)\, dx$$

$$= \exp(-k_1) + 1 - \exp(-k_2)$$

E3) The likelihood function is given by

$$L(p; \underline{X}) = \prod_{j=1}^{n} P(X = X_j) = p^{\sum_1^n X_j} (1-p)^{n - \sum_1^n X_j}$$

Let $r = \sum_1^n X_j$.

Now

$$\sup_{\underline{\theta} \in \Omega} L(p, \underline{X}) = \sup_{0 \leq p \leq 1} p^r (1-p)^{n-r}$$

The maximum likelihood estimate of p is $\hat{p} = r/n$. Thus

$$\sup_{0 \leq p \leq 1} p^r (1-p)^{n-r} = \left(\frac{r}{n}\right)^r \left(1 - \frac{r}{n}\right)^{n-r}$$

Also

$$\sup_{\theta \in \Omega_0} p^r (1-p)^{n-r} = \sup_{p \leq p_0} p^r (1-p)^{n-r}$$

The maximum likelihood estimate of p is p_0 if $p_0 < \frac{r}{n}$, and is $\frac{r}{n}$ if $p_0 \geq \frac{r}{n}$.

Thus,

$$\sup_{p \leq p_0} p^r (1-p)^{n-r} = p_0^r (1-p_0)^{n-r} \quad \text{if} \quad p_0 < r/n$$

$$= \left(\frac{r}{n}\right)^r \left(1 - \frac{r}{n}\right)^{n-r} \quad \text{if } p_0 \geq r/n$$

Now $\lambda(X) = \dfrac{p_0^r (1-p_0)^{n-r}}{\left(\frac{r}{n}\right)^r \left(1 - \frac{r}{n}\right)^{n-r}}$ if $p_0 < r/n$

$= 1$ if $p_0 \geq r/n$.

Since $\lambda(X) \leq 1$ for $r > np_0$ and $\lambda(X) + 1$ for $r \leq np_0$, $\lambda(X)$ is a decreasing function of r. Thus the test statistic $\lambda(X) < c$ implies $r > c'$, where c' is so choosen that

$$\sup_{p \leq p_0} P(r > c') = \alpha.$$

The distribution of r is binomial with parameters n and p, b (n, p) and

$$P_p(r > c') = \sum_{j=c'+1}^{n} \binom{n}{j} p^j (1-p)^{n-j}.$$

It can be seen that $P_p(r > c')$ is a non-decreasing function of p, so that

$$\sup_{p \le p_0} P(r > c') = \sum_{j=c'+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

Thus for a preassigned α, $0 < \alpha < 1$, choose c' so that

$$\alpha = \sum_{j=c'+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

Since r has a discrete distribution no c' may exist for which we get the exact probability α. In this case, choose c' such that

$$\alpha \ge \sum_{c'+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

$$\alpha < \sum_{c'}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}.$$

UNIT 18 COMMON TESTS AND CONFIDENCE INTERVALS

Structure

18.1 INTRODUCTION

In Unit 17, you have been introduced to the problem of testing of hypothesis and also to some basic concepts of the theory of testing of hypothesis. There you have studied two important procedures for testing statistical hypotheses, viz. using Neyman-Pearson Lemma and the likelihood ratio test. In this unit, you will be exposed to the problem of testing statistical hypotheses involving the parameters of some important distributions through some selected examples. In this unit, you will also be exposed to the problem of constructing confidence intervals for parameters of some important distributions through some selected examples. You will also learn the use of chi-square test for goodness of fit.

Objectives

After reading this unit, you should be able to:

- derive test statistic for various testing of hypotheses problems as well as to derive power functions,
- construct confidence intervals for parameters of various distributions,
- conduct large sample tests.

18.2 SOME COMMON TESTS OF HYPOTHESIS FOR NORMAL POPULATIONS

In Unit 17, we have already described with examples two procedures for testing statistical hypotheses. (See Example 1 and Example 2). In this section we will employ Neyman-Pearson Lamma and likelihood ratio test for testing of hypothesis related to a normal population.

Example 1: Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. It is desired to obtain a test statistic for testing $H_0 : \mu_1 = \mu_2$ against $\mu_1 : \mu_1 \neq \mu_2$ when $\sigma^2 (> 0)$ is unknown.

In order to obtain the test statistic, we use the likelihood ratio test. We have

$$\Omega = \left\{ (\mu_1, \mu_2, \sigma^2) : -\infty < \mu_1, \mu_2 < \infty, \sigma^2 > 0 \right\}$$

$$\Omega_0 = \left\{ \mu_1 = \mu_2 = \mu \text{ (say)}, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0 \right\}$$

We shall write $\theta = (\mu_1, \mu_2, \sigma^2)$

We have

$$\{\sup_{\theta \in \Omega_0} L(\underline{\theta} | X, Y)$$

$$= \sup \frac{1}{(2\pi)^{\frac{m+n}{2}} (\sigma^2)^{\frac{m+n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_1^m (X_i - \mu_1)^2 + \sum_1^n (Y_i - \mu_2)^2 \right] \right\}$$

Under H_0, $\mu_1 = \mu_2 = \mu$ and the maximum likelihood estimate of μ is $\hat{\mu} = \frac{m\overline{X} + n\overline{Y}}{m+n}$ and of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{m+n} \left[\sum_1^m (X_i - \mu_1)^2 + \sum_1^n (X_i - \mu_2)^2 + \frac{mn}{(m+n)} (\overline{X} - \overline{Y})^2 \right]$$

$= u'$ (say)

Thus $$\sup_{\theta \in \Omega_0} L(\underline{\theta} | X, Y) = \frac{1}{(2\pi u')^{\frac{m+n}{2}}} \exp\left[-\frac{1}{2u'} (m+n) u' \right]$$

$$= \left(\frac{1}{2\pi u'} \right)^{\frac{m+n}{2}} \exp\left(-\frac{(m+n)}{2} \right)$$

Under H_1, the maximum likelihood estimates of μ_1, μ_2 and σ^2 are respectively

$$\hat{\mu}_1 = \overline{X}, \hat{\mu}_2 = \overline{Y}, \hat{\sigma}^2 = \frac{\sum_1^m (X_i - \overline{X})^2 + \sum_i^n (Y_i - \overline{Y})^2}{m+n} = u \text{ (say)}$$

and

$$\sup_{\underline{\theta} \in \Omega} L(\theta | X, Y)$$

$$= \left(\frac{1}{2\pi u} \right)^{\frac{m+n}{2}} \exp\left(-\frac{m+n}{2} \right)$$

The likelihood ratio test is thus

$$\lambda(X, Y) = \frac{\sup_{\underline{\theta} \in \Omega_0} L(\underline{\theta} | X, Y)}{\sup_{\underline{\theta} \in \Omega} L(\underline{\theta} | X, Y)}$$

$$= \left(\frac{u}{u'} \right)^{\left(\frac{(m+n)}{2} \right)}$$

$$= \left[\frac{\sum_1^m (X_i - \overline{X})^2 + \sum_i^n (Y_i - \overline{Y})^2}{\sum_1^m (X_i - \overline{X})^2 + \sum_i^n (Y_i - \overline{Y})^2 + \frac{mn}{(m+n)} (\overline{X} - \overline{Y})^2} \right]^{\frac{m+n}{2}}$$

$$= \left[\frac{1}{1 + \dfrac{mn(\bar{X} - \bar{Y})^2}{(m+n)\left\{ \sum_1^m (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2 \right\}}} \right]$$

Now under null hypothesis, $\mu_1 = \mu_2 = \mu$, and $t = \dfrac{\bar{X} - \bar{Y}}{S\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}}$ follows a Student's t distribution with $m + n - 2$ degrees of freedom, where $S^2 = \dfrac{u(m+n)}{m+n-2}$

Thus

$$t^2 = \frac{(m+n-2)\, mn(\bar{X} - \bar{Y})^2}{(m+n)\left\{ \sum_1^m (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2 \right\}}$$

and

$$\lambda(X, Y) = \left[\frac{1}{1 + \dfrac{t^2}{m+n-2}} \right]^{\frac{m+n}{2}}$$

The likelihood ratio critical region is given by

$$\lambda(X, Y) = \left[\frac{1}{1 + \dfrac{t^2}{m+n-2}} \right]^{\frac{m+n}{2}} < c$$

$$\Rightarrow \quad \frac{1}{1 + \dfrac{t^2}{m+n-2}} < c^{\left(\frac{2}{m+n}\right)}$$

where c is to be determined so that

$$\sup_{\theta \in \Omega_0} P_\theta [\lambda(X, Y) < c] = \alpha$$

Since $\lambda(X, Y)$ is a decreasing function of $t^2/(m+n-2)$ we reject H_0 if

$$\frac{t^2}{(m+n-2)} > c^{2/(m+n)}$$

or

$$|t| > c_1$$

where c_1 is so chosen that

$$\sup_{\theta \in \Omega_0} P_\theta [\,|t| > c_1\,] = \alpha$$

Let $c_1 = t_{m+n-2,\,\alpha/2}$ in accordance with the distribution of t under H_0. Thus, the two sided test obtained is

$$\left| \frac{(\bar{X} - \bar{Y})}{S} \sqrt{\frac{mn}{(m+n)}} \right| > t_{m+n-2,\,\alpha/2}$$

Example 2: Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, μ is known and $\sigma^2 > 0$, is unknown. We wish to obtain a test statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ against an alternative $H_1 : \sigma^2 = \sigma_1^2 (> \sigma_0^2)$.

We have

$$P_{\theta_1}(\underline{X}) = \frac{1}{(2\pi\sigma_1)^{n/2}} \exp\left[-\frac{1}{2\sigma_1^2} \sum_1^n (X_i - \mu)^2 \right]$$

$$P_{\theta_0}(\underline{X}) = \frac{1}{(2\pi\sigma_0)^{n/2}} \exp\left[-\frac{1}{2\sigma_0^2} \sum_1^n (X_i - \mu)^2 \right]$$

Using Neyman-Pearson Lemma, the test statistic is

$$T(X) = \frac{P_{\theta_1}(\underline{X})}{P_{\theta_0}(\underline{X})} \geq k$$

$$\Rightarrow \left(\frac{\sigma_0}{\sigma_1} \right)^{n/2} \exp\left\{ 1/2 \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \sum_1^n (X_i - \mu)^2 \right.$$

$$\Rightarrow (\sigma_1^2 - \sigma_0^2) \sum_1^n (X_i - \mu^2) \geq k', \quad \text{taking logarithms}$$

$$\Rightarrow \sum_1^n (X_i - \mu^2) \geq k_1, \qquad \text{since } \sigma_1^2 > \sigma_0^2, \text{ under } H_1$$

Here k_1 is so determined that

$$P_{\theta_0}(T(\underline{X}) \geq k) = \alpha$$

$$\Rightarrow P_{\theta_0}\left[\sum_1^n (X_i - \mu)^2 \geq k_1 \right] = \alpha$$

$$\Rightarrow P_{\theta_0}\left[\sum_1^n (X_i - \mu)^2 / \sigma_0^2 \geq k_1 / \sigma_0^2 \right] = \alpha$$

Under the null hypothesis, since $\sigma^2 = \sigma_0^2$, $\sum_1^n (X_i - \mu)^2/\sigma_0^2$ has a χ_n^2 distribution (chi-square distribution with n degrees of freedom). Let $\chi^2_{n,\alpha}$ be the upper -α probability point of χ_n^2. The test statistic is thus

$$\sum_1^n (X_i - \mu)^2 > = k_1 \text{ and hence}$$

$$C_0 = \left\{ X \mid \sum_1^n (X_i - \mu)^2/\sigma_0^2 > \chi^2_{n,\alpha} \right\}$$

and $k_1/\sigma_0^2 = \chi^2_{n,\alpha} \Rightarrow k_1 = \sigma_0^2 \chi^2_{n,\alpha}$.

On the other hand, if the alternative hypothesis is $H_1 : \sigma^2 = \sigma_1^2 (\sigma_1^2 < \sigma_0^2)$, then the test statistic is

$$\sum_1^n (X_i - \mu)^2 < k_2$$

and hence

$$C_0 = \left\{ X \mid \Sigma (X_i - \mu)^2/\sigma^2 < \chi^2_{n,1-\alpha} \right\}$$

where $\chi^2_{n,1-\alpha}$ is the lower α -probability point of the χ^2 distribution with n degrees of freedom.

Example 3 : Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. We wish to obtain a test statistic for testing $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Here $\Omega = \left\{ (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, \sigma_1^2 > 0, i = 1, 2 \right\}$

and $\Omega_0 = \left\{ (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, i = 1, 2, \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0 \right\}$

We shall use $\underline{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

Also $L(\underline{\theta} \mid X, Y)$

$$= \left(\frac{1}{2\pi}\right)^{\frac{m+n}{2}} \left(\frac{1}{\sigma_1^2}\right)^{m/2} \left(\frac{1}{\sigma_1^2}\right)^{n/2} \exp\left\{ -\frac{1}{2\sigma_1^2} \sum_1^n (X_i - \mu_1)^2 - \frac{1}{2\sigma_2^2} \sum_1^n (Y_i - \mu_2)^2 \right\}$$

The maximum likelihood estimates of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are respectively

$$\hat{\mu}_1 = \frac{1}{m} \sum_1^m X_i = \bar{X}, \quad \hat{\mu}_2 = \frac{1}{n} \sum_1^n Y_i = \bar{Y}$$

$$\hat{\sigma}_1^2 = \frac{1}{m}\sum_1^m (X_i - \bar{X})^2, \hat{\sigma}_2^2 = \frac{1}{n}\sum_1^n (Y_i - \bar{Y})^2$$

Further, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{(m+n)}\left[\sum_1^m (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2\right]$$

Thus

$$\sup_{\underline{\theta} \in \Omega_0} L(\theta \mid X, Y)$$

$$= \frac{\exp\{-(m+n)/2\}}{[2\pi/(m+n)]^{\frac{m+n}{2}} \left\{\sum_1^m (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2\right\}^{\frac{m+n}{2}}}$$

and

$$\sup_{\underline{\theta} \in \Omega_0} L(\theta \mid X, Y)$$

$$= \frac{\exp\{-(m+n)/2\}}{(2\pi/m)^{m/2}(2\pi/n)^{n/2}\left\{\sum_1^m (X_i - \bar{X})^2\right\}^{\frac{m}{2}}\left\{\sum_1^n (Y_i - \bar{Y})^2\right\}^{\frac{n}{2}}}$$

The likelihood ratio test is thus

$$\lambda(X, Y) = \frac{\sup_{\underline{\theta} \in \Omega_0} L(\underline{\theta} \mid X, Y)}{\sup_{\underline{\theta} \in \Omega} L(\underline{\theta} \mid X, Y)}$$

$$= \left(\frac{m}{m+n}\right)^{m/2}\left(\frac{n}{m+n}\right)^{n/2} \frac{\left\{\sum_1^m (X_i - \bar{X})^2\right\}^{\frac{m}{2}}\left\{\sum_1^n (Y_i - \bar{Y})^2\right\}^{\frac{n}{2}}}{\left\{\sum_1^m (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2\right\}^{\frac{m+n}{2}}}$$

Now

$$\frac{\left\{\sum_1^m (X_i - \bar{X})^2\right\}^{\frac{m}{2}}\left\{\sum_i^n (Y_i - \bar{Y})^2\right\}^{\frac{n}{2}}}{\left\{\sum_1^m (X_i - \bar{X})^2 + \sum_1^n (Y_i - \bar{Y})^2\right\}^{\frac{m+n}{2}}}$$

$$\frac{1}{\left[1+\sum_1^n (Y_i-\bar{Y})^2/\sum_1^m (X_i-\bar{X})^2\right]^{\frac{m}{2}}\left[1+\sum_1^m (X_i-\bar{X})^2/\sum_1^n (Y_i-\bar{Y})^2\right]^{\frac{n}{2}}}$$

Writing $f = \dfrac{\sum_1^m (X_i-\bar{X})^2/(m-1)}{\sum_1^n (Y_i-\bar{Y})^2/(n-1)}$

we have

$$\lambda(X, Y) = \frac{\left(\frac{m}{m+n}\right)^{m/2}\left(\frac{n}{m+n}\right)^{m/2}}{\left[1+\frac{(m-1)}{(n-1)}f\right]^{n/2}\left[1+\frac{(n-1)}{(m-1)}(1/f)\right]^{m/2}}$$

The likelihood ratio test criterion rejects H_0 if $\lambda(X, Y) < c$

It is easy to see that $\lambda(X, Y)$ is a monotonic function of f and $\lambda(X, Y) < c$ is equivalent to $f < c_1$ or $f > c_2$. Under H_0,

$$f = \frac{\sum_1^m (X_i-\bar{X})^2/(m-1)}{\sum_1^n (Y_i-\bar{Y})^2/(n-1)}$$

has Snedecor's $F(m-1, n-1)$ distribution, so that c_1, c_2 can be selected, such that

$$\sup_{\theta \in \Omega_0} P_\theta[\lambda(X, Y) < c] = \alpha$$

or

$$P(F \le c_1) = P(F \ge c_2) = \alpha/2$$

Thus $c_2 = F(m-1, n-1, \alpha/2)$ is the upper $\alpha/2$ probability point of $F(m-1, n-1)$ distribution and $c_1 = F(m-1, n-1, 1-\alpha/2)$ is the lower probability point of $F(m-1, n-1)$

E1) Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, both μ and σ^2 unknown. Obtain a test statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ against all its alternatives.

E2) Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$. Obtain a test for testing $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$, say.

18.3 CONFIDENCE INTERVALS

In Section 15.6 of Unit 15 you have been briefly exposed to some notions of interval estimation of a parameter. In this section we discuss in detail the problem of obtaining interval estimates of parameters and describe, through examples, some methods of constructing interval extimates of parameters. We may remind you again that an interval estimate is also called a confidence interval or a confidence set. We first illustrate through small examples the need for constructing confidence intervals. Suppose X denotes the tensile strength of a copper wire. A potential user may desire to know the lower bound for the mean of X, so that he can use the wire if the average tensile strength is not less than say θ_0. Similarly, if the random variable X measures the toxicity of a drug, a doctor may wish to have a knowledge of the upper bound for the mean of X in order to prescribe this drug. If the random variable X measures the waiting times at the emergency room of a large city hospital, one may be interested in the mean waiting time at this emergency room. In this case we wish to obtain both the lower and upper bounds for the waiting time.

In this unit we are concerned with the problem of determining confidence intervals for a parameter. A formal definition of a confidence interval has been given in Section 15.6. However, for the sake of completeness we define some terms here.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density (or, mass) function $f(x, \theta)$, $\theta \in \Omega \subseteq R^1$. The object is to find statistics $r_L(X_1 \ldots\ldots, X_n)$ and $r_U(X_1, \ldots, X_n)$ such that

$P_\theta\{(r_L(X_1, \ldots, X_n) \le \theta \le r_U(X_1, \ldots, X_n)\} \ge 1-\alpha$ for all $\theta \in \Omega \subseteq R^1$. The interval $\left(r_L(\underline{X}), r_U(\underline{X})\right)$ is called a confidence interval and the quantity

$$\inf P_\theta\left[r_L(X_1, \ldots, X_n) \le \theta \le r_U(X_1, \ldots, X_n)\right]$$

will be referred to as the confidence co-efficient associated with the random interval.

We now give some examples of construction of confidence intervals.

Example 4: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population, $N(\mu, \sigma^2)$. We wish to obtain a $(1-\alpha)$ level confidence interval for μ.

Let $\overline{X} = n^{-1}\sum_{1}^{n} X_i$. Consider the interval $(\overline{X}-a, \overline{X}+b)$. In order for this to be a $(1-\alpha)$ level confidence interval, we must have

$$P\{\overline{X}-a < \mu < \overline{X}+b\} \ge 1-\alpha$$

Thus

$$P\left\{-\frac{b}{\sigma}\sqrt{n} < \frac{(\overline{X}-\mu)}{\sigma}\sqrt{n} < \frac{a}{b}\sqrt{n}\right\} \ge 1-\alpha$$

Since, $\frac{(\overline{X}-\mu)}{\sigma}\sqrt{n} \sim N(0, 1)$ we can choose a and b to satisfy

$$P\left\{-\frac{b}{\sigma}\sqrt{n} < \frac{(\overline{X}-\mu)\sqrt{n}}{\sigma} < \frac{a}{\sigma}\sqrt{n}\right\} = 1-\alpha$$

provided that σ is known. There are infinitely many such pairs of values (a, b). In particular, an intuitively reasonable choice is $a = b = c$, say

In that case

$\frac{c\sqrt{n}}{\sigma} = Z_{\alpha/2}$ where $Z_{\alpha/2}$ is the $\alpha/2$ percent point of the standard normal distribution, and the confidence interval is

$$(\bar{X} - (\sigma/\sqrt{n})\, Z_{\alpha/2}, \bar{X} + (\sigma/\sqrt{n})\, Z_{\alpha/2})$$

The length of the interval is $(2\sigma/\sqrt{n})\, Z_{\alpha/2}$ Given σ and α one can choose n to get a confidence interval of desired length.

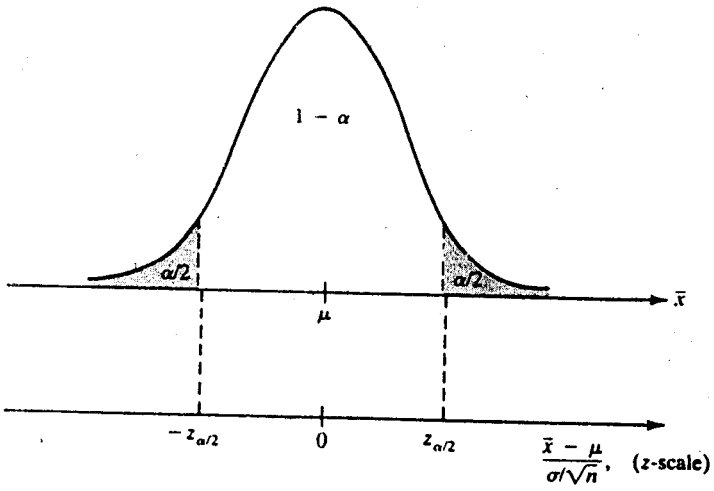


Figure 1 : Probability density curve of normal distribution with mean μ and variaance σ^2/n. Shows area $\alpha/2$ in each of two tails

If σ^2 is unknown, we have from

$$P\{-b < \bar{X} - \mu < a\} \geq 1 - \alpha$$

that

$$P\left\{-\frac{b}{S}\sqrt{n} < \frac{(\bar{X} - \mu)}{S} < \frac{a}{S}\sqrt{n}\right\} \geq 1 - \alpha$$

and $$S^2 = (n-1)^{-1} \sum_{1}^{n} (X_i - \bar{X})^2.$$

It is known that $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$. We can choose pairs of values (a, b) using a students *t*-distribution with (n – 1) degrees of freedom such that

$$P\left\{-\frac{b\sqrt{n}}{S} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < \frac{a\sqrt{n}}{S}\right\} = 1 - \alpha$$

In particular, an intuitively reasonable choice is $a = b = c$ say. Then

$$\frac{c\sqrt{n}}{S} = t_{n-1,\,\alpha/2}$$

and $\left(\bar{X} - (S/\sqrt{n})\, t_{n-1,\alpha/2},\ \bar{X} + (S/\sqrt{n})\, t_{n-1,\alpha/2}\right)$ is $1-\alpha$ level confidence interval for μ. The length of the interval is $(2S/\sqrt{n})\, t_{n-1,\alpha/2}$, which is no longer constant. Therefore, in this case one cannot choose n to get a fixed length confidence interval of level $1-\alpha$. The expected length is, however,

$$\frac{2}{\sqrt{n}}\, t_{n-1,\alpha/2}\, E_\sigma(S) = \frac{2}{\sqrt{n}}\, t_{n-1,\alpha/2} \sqrt{\frac{2}{n-1}}\, \frac{\Gamma(n/2)}{\Gamma(n-1)/2)}\, \sigma$$

which can be made as small as we want by making a proper choice of n for a given σ and α.

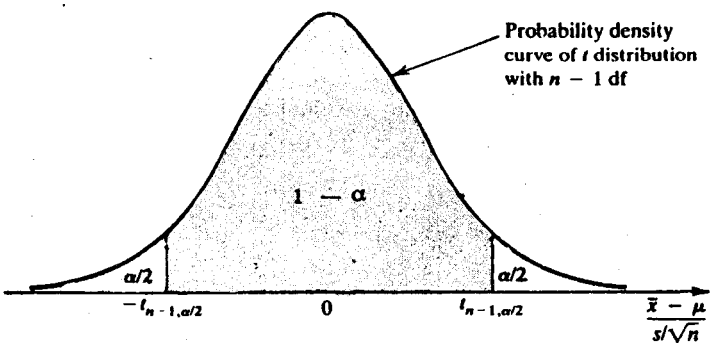


Figure 2 : t **values such that there is an area** $\alpha/2$ **in the right tail and** $\alpha/2$ **in the left tail of the distribution.**

Example 5 : Let $X_1, X_2, \ldots, X_n$ be a random sample, from $N(\mu, \sigma^2)$. It is desired to obtain a confidence interval for σ^2 when μ is unknown.

Consider the interval (aS^2, bS^2), $a, b > 0$, $S^2 = (n-1)^{-1} \sum_1^n (X_i - \bar{X})^2$. We have

$$P\left\{aS^2 < \sigma^2 < bS^2\right\} \geq 1-\alpha$$

o that

$$P\left\{b^{-1} < \frac{S^2}{\sigma^2} < a^{-1}\right\} \geq 1-\alpha$$

is known that

$$(n-1)\, S^2/\sigma^2 \sim \chi^2_{n-1}.$$

Ve can therefore choose pairs of intervals (a, b) from the tables of the chi-square listribution. In particular we can choose a, b so that

$$P\left\{\frac{S^2}{\sigma^2} \geq \frac{1}{a}\right\} = \alpha/2 = P\left\{\frac{S^2}{\sigma^2} \leq \frac{1}{b}\right\}.$$

Then $\frac{n-1}{a} = \chi^2_{n-1,\,\alpha/2}$ and $\frac{n-1}{b} = \chi^2_{n-1,\,1-\alpha/2}$ and the $1-\alpha$ level confidence interval for σ^2 when μ is unknown is

$$\left(\frac{(n-1)\,S^2}{\chi^2_{n-1,\,\alpha/2}},\ \frac{(n-1)S^2}{\chi^2_{n-1,\,1-\alpha/2}} \right)$$

If however, μ is known then $(n-1)\,S^2$ is replaced by $\sum_1^n (X_i-\mu)^2$ and the degrees of freedom of χ^2 is n instead of $n-1$, for $\sum_1^n (X_i-\mu)^2/\sigma^2 \sim \chi^2_n$.

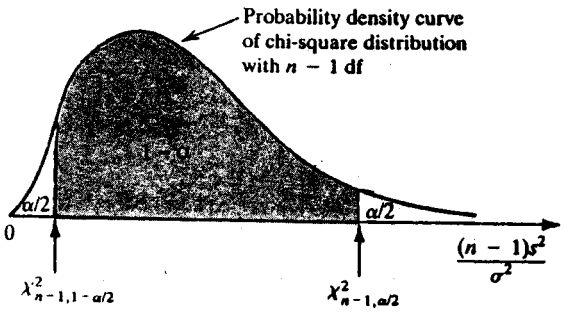


Figure 3 : Chi-square values such that area $1-\alpha/2$ and $\alpha/2$ are to their right.

Example 6: Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ denote respectively independent random samples from the two independent distributions having respectively the probability density functions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$. We wish to obtain a confidence interval for $\mu_1 - \mu_2$.

Consider the interval $\left\{ (\overline{X}-\overline{Y})-a,\ (\overline{X}-\overline{Y})+b \right\}$. In order that this is a $(1-\alpha)$ level confidence interval, we must have

$$P\left\{ (\overline{X}-\overline{Y})-a < \mu_1-\mu_2 < (\overline{X}-\overline{Y})+b \right\} \geq 1-\alpha$$

which is the same as

$$P\left\{ -b < (\overline{X}-\overline{Y})-(\mu_1-\mu_2) < a \right\} \geq 1-\alpha$$

or

$$P\left\{ \frac{-b}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{a}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} \right\} \geq 1-\alpha$$

64

Here $\bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$ and $\bar{Y} = \frac{1}{m}\sum_{1}^{m} Y_i$

Since $$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim N(0,1).$$

we can choose a and b to satisfy

$$P\left\{\frac{-b}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{a}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\right\} = 1-\alpha$$

provided that σ is known . There are infinitely many such pairs of values (a, b). In particular, an intuitively reasonable choice is a = b = c, say. In that case

$c/\left\{\sigma\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2}\right\} = Z_{\alpha/2}$ and the confidence interval is

$$\left\{(\bar{X}-\bar{Y})-\sigma\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} Z_{\alpha/2}, (\bar{X}-\bar{Y})+\sigma\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} Z_{\alpha/2}\right\}$$

The length of the interval is $2\sigma\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} Z_{\alpha/2}$. Given α and σ one can choose n and m to get a desired length confidence interval.

If σ^2 is unknown, we have from

$$P\{-b < (\bar{X}-\bar{Y})-(\mu_1-\mu_2) < a\} \geq 1-\alpha$$

that

$$P\left\{\frac{-b}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{a}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\right\} \geq 1-\alpha$$

where $$S^2 = \frac{\sum_{1}^{n}(X_i-\bar{X})^2 + \sum_{1}^{m}(Y_i-\bar{Y})^2}{(n+m-2)} = \frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}$$

It is known that

$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S^2\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim t_{n+m-2}$. We can choose pairs of values (a, b) using Student's t-distribution with n + m – 2 degrees of freedom such that

$$P\left\{\frac{-b}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{a}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\right\} = 1-\alpha$$

In particular, an intuitively reasonable choice is a = b = c, say. Then

$$\frac{c}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} = t_{n+m-2,\,\alpha/2},$$

and $\left\{(\bar{X}-\bar{Y})-S\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} t_{n+m-2,\,\alpha/2},\ (\bar{X}-\bar{Y})+S\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} t_{n+m-2,\,\alpha/2}\right\}$

is a $1-\alpha$ level confidence interval for $\mu_1-\mu_2$.

Example 7: Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, n, m > 2, denote respectively independent random samples from the two distributions having respectively the probability density functions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. We wish to obtain a confidence interval for the ratio σ_2^2/σ_1^2 when μ_1 and μ_2 are unknown.

Consider the interval $\left(a\, S_2^2/S_1^2,\ b\, S_2^2/S_1^2\right)$ a, b > 0, where

$$S_1^2 = \frac{1}{(n-1)}\sum_1^n (X_i-\bar{X})^2,\quad S_2^2 = \frac{1}{(m-1)}\sum_1^m (Y_i-\bar{Y})^2,$$

$$\bar{X} = \frac{1}{n}\sum_1^n X_i,\quad \bar{Y} = \frac{1}{m}\sum_1^m Y_i.$$ We have

$$P\left\{a\frac{S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < b\frac{S_2^2}{S_1^2}\right\} \geq 1-\alpha$$

so that

$$P\left\{\frac{1}{b} < \frac{\left(S_2^2/S_1^2\right)}{\left(\sigma_2^2/\sigma_1^2\right)} < \frac{1}{a}\right\} \geq 1-\alpha$$

It is known that $(n-1)\,S_1^2/\sigma_1^2 \sim \chi^2_{(n-1)}$ and $(m-1)\,S_2^2/\sigma_2^2 \sim \chi^2_{(m-1)}$.

It is also known that if X and Y are independent χ^2 random variables with m and n degrees of freedom respectively, the random variable $F = (X/m)/(Y/n)$ is said to have an F-distribution with (m, n) degrees of freedom. It is also known that if X has an F (m, n) distribution then 1/X has an F (n, m) distribution, and $F_{m,n,1-\alpha} = 1/F_{n,m,\alpha}$. Therefore

$$\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} = \frac{S_2^2/S_1^2}{\sigma_2^2/\sigma_1^2} \sim F_{(m-1),(n-1)}$$

We can therefore choose paris of values (a, b) from the tables of F-distribution. In particular, we can choose a and b so that

$$P\left\{\frac{(S_2^2/\sigma_2^2)}{(S_1^2/\sigma_1^2)} \geq \frac{1}{a}\right\} = \alpha/2 = P\left\{\frac{(S_2^2/\sigma_2^2)}{(S_1^2/\sigma_1^2)} \leq \frac{1}{b}\right\}.$$

Then $\frac{1}{a} = F_{m, n, \alpha/2}$ and $\frac{1}{b} = F_{m, n, 1-\alpha/2}$ and the $1 - \alpha$ level confidence interval for σ_2^2/σ_1^2 is

$$\left(\frac{S_2^2}{S_1^2} - F_{n, m, 1-\alpha/2}, \frac{S_2^2}{S_1^2} - F_{n, m, 1/\alpha/2}\right)$$

E3) Let $X_1, \ldots, X_n$ be independently and identically distributed b (1, p) random variables. Obtain a confidence interval for p.

18.4 CHI SQUARE TEST FOR GOODNESS OF FIT

An important limiting distribution used in connection with categorical data is the χ^2-distribution. By categorical data we mean data which are presented in the form of frequencies falling in different categories or classes. The categorisation may be with respect to a character which is either an attribute or a variable. The categorisation may also be with respect to two or more characters.

In this section we introduce some tests of statistical hypothesis that are commonly called χ^2-tests. These tests have been immensely useful in practical applications particularly in problems dealing with categorical data. A test of this sort was originally proposed by Karl Pearson before the formal theory of testing of hypothesis was developed.

Suppose $X_1, \ldots, X_n$ is a random sample from a discrete distribution given by

$$P[X = x_j] = p_j, \; j = 1, \ldots, k$$

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

where $p_i > 0$ and $\sum_1^k p_j = 1$. We wish to obtain a test statistic for testing the null hypothesis

$$H_0 : p_j = p_{0j}, \; j = 1, \ldots, k,$$

where p_{0j} are known numbers, on the basis of the observations.

Let n_j be the number of X's that equal X_j in a sample of size n. Then

n_j's, j = i, . . . , k are random variables, satisfying $\sum_1^k n_j = n$. Further, the distribution of $(n_1, \ldots, n_{k-1})$ is under H_0, multinominal with parameters

$o_1, \ldots, P_{0(k-1)})$. We are not considering n_k (p_k) because $\sum_1^k n_j$ and $\sum_1^n p_j$ are fixed.

It is known that n_j has a marginal binomial distribution with parameter p_j Hence

$$E(n_j) = np_j, \text{ so that } E(n_j/n) = p_j$$

and we can use n_j/n, $j = 1, \ldots, k$ respectively as unbiased estimates of $p_1, \ldots, p_k$. Intuitively, it appears meaningful to compare the observed proportions n_j/n with the postulated proportions p_{0j}, $j = 1, \ldots, k$.. The various way which can be used for such a comparisonare, e.g.,

$$\sum_1^k \left| n_j/n - p_{0j} \right| \text{ or } \sum_1^k \left(n_j/n - p_{0j} \right)^2$$

Another measure most commonly used in practice is the weighted sum of squares

$$Q = \sum_1^k \left(\frac{n}{p_{0j}} \right) \left(n_j/n - p_{0j} \right)^2$$

$$= \sum_1^k \frac{\left(n_j - np_{0j} \right)^2}{np_{0j}}$$

Since, under H_0, np_{0j}, is the expected value of n_j, we would feel intuitively that the observed value of Q should not be large if H_0 is true. Thus large values of Q would lead to large discrepancy between the observed data and the postulated hypothesis and thus lead to the rejection of H_0.

What is the distribution of Q under the null hypothesis? This is answered below. We however, omit the proof,

Theorem 2 : In a random and large sample of size n,

$$Q = \sum_{j=1}^k \frac{(n_j - np_{0j})^2}{np_{0j}}$$

follows approximately chi-square distribution with k-1 degrees of freedom, where n_j is the observed frequency and np_{0j} is the corresponding expected frequency of the j^{th} class ($j = 1, \ldots, k$), $\sum_1^k n_j = n$..

Since the distribution of the random variable Q, under H_0 is approximately χ^2 with k-1 degrees freedom, the test rejects the null hypothesis H_0 if $Q \geq c$ where c is so determined that

$$P_{H_0}(Q \geq c) = \alpha \text{ (–the desired level of significance).}$$

Thus, we get $c = \chi^2_{k-1, \alpha}$ and reject the hypothesis H_0 at level α if $Q > x^2_{k-1, \alpha}$.

For $k = 2$, Q has approximately a chi-square distribution with 1 degree of freedom. Indeed, for $k = 2$, the problem reduces to testing $H_0 : p = p_0$ against $H : p \neq p_0$ in sampling from binomial population.

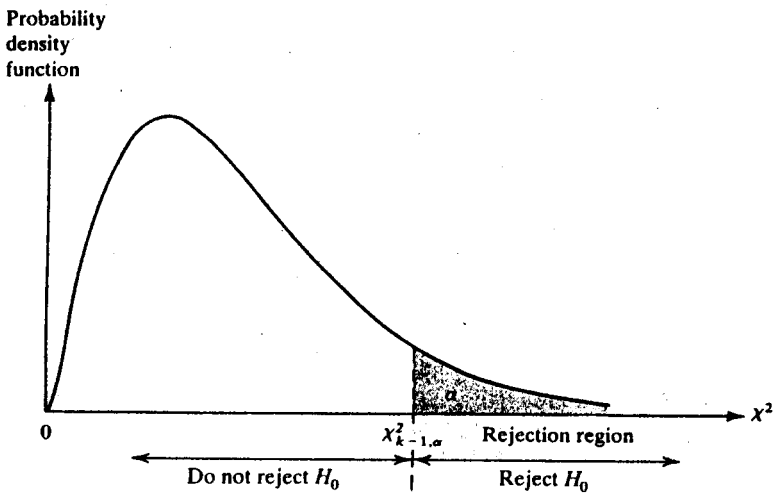


Figure 4 : Probability density curve for chi-square distribution with k – 1 df and the decision rule.

Also under H_0,

$$E(Q) = \sum_1^k E\left\{\frac{(n_j - np_{0j})^2}{np_{0j}}\right\} = \sum_1^k \frac{1}{np_{0j}} \text{var}(n_j)$$

$$= \sum_1^k \frac{1}{np_{0j}} np_{0j}(1-p_0) = \sum_1^k (1-p_{0j})$$

$$= k-1.$$

We can use the test statistic Q to test that a random variable X has a specified probability density function f. In order to do that, we divide the entire real line $(-\infty, \infty)$ into k intervals $I_1, \ldots, I_k$ and let $p_{0j} = p_f[X \in I_j], j = 1, \ldots, k$. Let $\hat{p}_j$ be the observed proportion of observations in the sample that belong to the interval I_j

Then the statistic

$$Q = \sum_1^k \frac{(n\hat{p}_j - np_{0j})^2}{np_{0j}}$$

has approximately a chi-square distribution with $k-1$ degrees of freedom, under H. But, then an important question which arises now is as to how to choose the intervals $I_1, \ldots, I_k$. The answer to this question comes from the fact that the chi-square distribution is only an approximation to the true distribution of Q. So the choice has to be made in such a way that this approximation is good. Secondly, the underlying assumption to the chi-square approximation is that each random variable $\frac{(n\hat{p}_j - np_{0j})}{\sqrt{np_{0J}}}$ has approximately a normal distribution with mean zero and variance $1 - p_{0j}$. This holds provided n is large .

But if np_{0j} is small, then the term $\left(n\hat{p}_j - jp_{0j}\right)^2/np_{0j}$ in Q will have a dominating effect on the other terms because of a small denominator. Although the approximation is often good even for np_{0j} as small as one the following rules of thumb are generally used in deciding upon the intervals $I_1, \ldots, I_k$.

i) Choose $I_1, \ldots, I_k$ such that under H_0, $p_{0j} = p[X \in I_j]$ is approximately equal to $1/k$ and each $np_{0j} = \frac{n}{k} \geq 5$.

ii) If any np_{0j}'s is less than 5, pool the corresponding interval with one or more intervals to make the cell frequency at least 5. The decision which intervals (or cells) to pool is arbitrary but we restrict pooling to a minimum. The degrees of freedom associated with the chi-square approximation, after pooling, are reduced by the number of classes pooled.

Example 8: A random sample of 80 points is picked from the interval (0,1) so that the underlying density is $f(x) = 1, 0 < x < 1$, and zero elsewhere. The data are as follows :

.03	.69	.30	.92	.10	.85	.53	.6
.35	.01	.70	.24	.22	.93	.22	
.10	.37	.25	.65	.36	.64	.95	
.81	.15	.41	.74	.66	.31	.06	
.18	.34	.38	.04	.99	.17	.91	
.09	.47	.13	.36	.54	.35	.45	
.70	.33	.77	.79	.13	.72	.52	
.29	.89	.60	.33	.38	.40	.72	
.91	.57	.28	.47	.11	.69	.14	

We construct the frequency distribution with equal class intervals of equal width. We wish to test the null hypothesis H_0, that the sample comes from the uniform distribution on (0,1).

The following table gives the observed and the expected frequencies.

Interval	(.005, .105)	(.105, .205)	(.205, .305)	(.305, .405)	(.405, 505)
n_j	8	11	8	13	6
np_{0j}	8	8	8	8	8
Interval	(.505, .605)	(.605, .705)	(.705, .805)	(.805, .905)	(.905, 1.00)
n_j	5	8	9	4	8
np_{0j}	8	8	8	8	8

We have $Q = \sum_{1}^{10} \frac{(n_j - np_{0j})^2}{np_{0j}} = 8.0$

The value of chi-square at 9 degrees of freedom and .05 level of significance is 16.919.

Hence the observed value of Q is not significant at 5 percent level and we cannot reject H_0.

We have just described the chi-square test of goodness of fit when the cell probabilities are prespecified. But a more interesting situation is one when instead of fitting the data reasonably with one specified distribution, the data can be fitted by one of a family of distributions. The question to be answered in such a situation is; could the observations have come from some Poission distribution ? Some normal distribution ? Some binomial distribution ? Some exponential distribution ? etc. In particular the null hypothesis H_0 is not a simple hypothesis now. Under H_0, the probability mass or density function is specified except for one or more parameters.

Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution given by

$P[X = x_j] = p_j(\theta), j = 1, \ldots, k$ and zero elsewhere, where

$p_j(\theta) > 0, \sum_1^k p_j(\theta) = 1$ and θ is a scalar or a vector parameter that is unknown. Let

n_j be the number of X's in the sample that equal x_j. Then $\sum_1^k n_j = n$ and under

$H_0 : (n_1, \ldots, n_{k-1})$ has a multinomial distribution . Let

$$Q(\theta) = \sum_1^k \frac{[n_j - np_j(\theta)]^2}{np_j(\theta)}$$

We wish to test the null hypothesis that for some value of θ, $p_j(\theta), j = 1, \ldots, k$, is a good fit to the data. To achieve this, minimise $Q(\theta)$ for all possible values of θ in the parameter space. In other words, find that value of θ for which $p_j(\theta)$ best fits the data. Let $\hat{\theta}$ be the value of θ for which $Q(\theta)$ is minimum. $\hat{\theta}$ is called the minimum chi-square estimate of θ. Under certain conditions and for large n the minimum value $Q(\hat{\theta})$ of $Q(\theta)$ given by

$$Q(\hat{\theta}) = \sum_1^k \frac{[n_j - np_j(\hat{\theta})]^2}{np_j(\hat{\theta})}$$

has approximately a chi-square distribution with $k - 1 - s$ degrees of freedom where s is the dimensionality of θ. If θ is scalar, then $s = 1$, if $\underline{\theta} = (\mu, \sigma^2)$, both μ and σ^2 unknown, then $s = 2$, and so on. It can be shown that for sufficiently large n the θ that minimizes $Q(\theta)$ is approximately the maximum likelihood estimate of θ and moreover $Q(\hat{\theta})$ has a chi-square distribution , when $\hat{\theta}$ is the maximum likelihood estimate of θ.

We reject H_0 at level α if $Q(\hat{\theta}) \geq \chi^2_{k-1,\alpha}$.

In the continuous case the procedure is the same as described earlier. let X be of continuous type with density function $f(x, \theta)$ where θ is unknown. Partition the real line into k intervals $I_1, \ldots, I_k$. under H_0,

$$p_j(\theta) = \int_{I_j} f(x, \theta)\, dx,$$

and the result stated above applies. We illustrate these ideas through an example.

Example 9: A random sample of 200 families, each with four children has the following frequency distribution of the number of girls.

Number of girls	:	0	1	2	3	4
Number of families	:	5	32	65	75	23

We wish to test if the data comes from a binomial distribution.

Let X denote the number of girls in any family of four children. We wish to test the composite null hypothesis

$$H_0 : P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n$$

against all alternatives. In this example n = 4. Since p is unknown, we compute the maximum likelihood estimate of p. We have

$$\hat{p} = 0.5987$$

In order to apply the test, we need

$$p_j(\hat{p}) = \binom{n}{j} \hat{p}^j (1-\hat{p})^{n-j}, j = 0, \ldots, n$$

We have

$$p_0(\hat{p}) = .02594,\ p_1(\hat{p}) = .15476,\ p_2(\hat{p}) = .34634$$

$$p_3(\hat{p}) = .34448,\ p_4(\hat{p}) = .12848$$

We now compare the observed and expected frequencies as follows :

j	:	0	1	2	3	4
Observed	:	5	32	65	75	23
Expected	:	5	31	69	69	26

Thus, we get

$$Q(\hat{p}) = \sum_1^k \frac{[n_j - np_j(\hat{p})]^2}{np_j(\hat{p})}$$

$$= 1.132$$

The number of degrees of freedom on the conservative side is 3. Since the value of $\chi^2_{3,0.05}$ is 7.81, there is not much evidence to reject H_0. We therefore infer that the data could have come from a binomial distribution.

18.5 SUMMARY

In this unit we have

1. demonstrated the use of Neyman-Pearson Lemma and likelihood ratio test for testing hypothesis about parameters of normal populations,
2. obtained confidence intervals for parameters of some standard distributions,
3. described the use of chi-square test for testing the goodness of fit problems.

18.6 SOLUTIONS AND ANSWERS

E1) We have

$$\Omega_o = \left\{(\mu, \sigma^2), -\infty < \mu < \infty, \sigma^2 = \sigma_0^2\right\}$$

and

$$\Omega = \left\{(\mu, \sigma^2), -\infty < \mu < \infty, \sigma^2 > 0\right\}$$

Now

$$\sup_{\underline{\theta} \in \Omega_o} L(\underline{\theta} | \underline{X}) = \sup_{\mu, \sigma^2 = \sigma_0^2} \frac{1}{(2\pi\sigma_0)^{n/2}} \exp\left[-\frac{1}{2\sigma_0^2} \sum_1^n (X_i - \mu)^2\right]$$

The maximum likelihood estimate of μ is

$$\hat{\mu} = 1/n \sum_1^n X_i = \bar{X}$$

Thus

$$\sup_{\underline{\theta} \in \Omega_0} L(\underline{\theta} | \underline{X}) = \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left[-\frac{1}{2\sigma_0^2} \sum_1^n (X_i - \bar{X})^2 \right]$$

Also

$$\sup_{\underline{\theta} \in \Omega} L(\underline{\theta} | \underline{X}) = \sup_{\mu, \sigma^2} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_1^n (X_i - \mu)^2 \right]$$

The maximum likelihood estimates of μ and σ^2 are respectively $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = 1/n \sum_1^n (x_i - \bar{x})^2$.

Thus

$$\sup_{\underline{\theta} \in \Omega} L(\underline{\theta} | \underline{X}) = \left[\frac{n}{2\pi \sum_1^n (X_i - \bar{X})^2} \right]^{n/2} \exp(-n/2)$$

The likelihood ratio test is

$$\lambda(X) = \frac{\sup_{\underline{\theta} \in \Omega_0} L(\underline{\theta} | \underline{X})}{\sup_{\underline{\theta} \in \Omega} L(\underline{\theta} | \underline{X})}$$

$$= \left[\frac{1/n \sum (X_i - \bar{X})^2}{\sigma_0^2} \right]^{n/2} \exp\left[-1/2 \left\{ -\frac{1}{\sigma_0^2} - \sum_1^n (X_i - \bar{X})^2 - n \right\} \right]$$

Under the null hypothesis

$$u = \frac{\sum_1^n (X_i - \bar{X})^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Thus

$$\lambda(\underline{X}) = \left(\frac{u}{n} \right)^{n/2} \exp\{-1/2(u - n)\}$$

It therefore follows that $\lambda(\underline{x}) < c$ is equivalent to $u < c_1$ or $u > c_2$. Since under H_0, u has a χ^2_{n-1} distribution, c_1 can c_2 can be selected. It is usual to take

$$P\{u \geq c_2\} = \alpha/2 \text{ and } P\{u \leq c_1\} = 1 - \alpha/2$$

In other words c_2 is the upper $\alpha/2$ probability point of χ^2_{n-1} and c_1 is the lower $1 - \alpha/2$ probability point of χ^2_{n-1}.

E2) We have

$$\Omega_0 = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) : -\infty < \mu_i < \infty, \sigma_i^2 > 0, i = 1, 2, \rho = 0\}$$

And

$$\Omega = \{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) : -\infty < \mu_i < \infty, \sigma_i^2 > 0, i = 1, 2, |\rho| < 1\}$$

Also we shall write $\underline{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

and

$$f(X, Y, \underline{\theta}) = \frac{1}{\left[[2\pi \sigma_1 \sigma_2 (1-\rho^2)^{1/2}]^n \right.}$$

$$\exp\left\{ -\frac{1}{2(1-\rho)^2} \sum_1^n \left[\left(\frac{X_i - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{X_i - \mu_1}{\sigma_1} \right)\left(\frac{Y_i - \mu_2}{\sigma_2} \right) + \left(\frac{Y_i - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

Under Ω_0, $\rho = 0$, and the maximum likelihood estimates of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are given respectively by

$$\hat{\mu}_1 = 1/n \sum_1^n X_i = \overline{X}, \quad \hat{\mu}_2 = \frac{1}{n} \sum_1^n Y_i = \overline{Y}$$

$$\hat{\sigma}_1^2 = \frac{1}{n} \sum_1^n (X_i - \overline{X})^2, \quad \hat{\sigma}_2^2 = \frac{1}{n} \sum_1^n (Y_i - \overline{Y})^2$$

Thus

$$\sup_{\underline{\theta} \in \Omega_0} L(\underline{\theta} \mid \underline{X}, \underline{Y}) = \frac{1}{(2\pi \hat{\sigma}_1 \hat{\sigma}_2)^n} \exp(-n)$$

under Ω, the maximum likelihood estimates of μ_1, μ_2, σ_1^2 and σ_2^2 are the same as under Ω_0, whereas the maximum likelihood estimate of ρ is

$$\hat{\rho} = r = \frac{\sum_1^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_1^n (X_i - \overline{X})^2 \sum_i^n (Y_i - \overline{Y})^2}}$$

$$= \frac{\sum_1^n (X_i - \overline{X})(Y_i - \overline{Y})}{n \hat{\sigma}_1 \hat{\sigma}_2}$$

Thus

$$\sup_{\theta \in \Omega_0} L(\theta \mid \underline{X}, \underline{Y}) = \frac{1}{(2\pi \hat{\sigma}_1 \hat{\sigma}_2)^n (1-r^2)^{n/2}} \exp(-n)$$

The likelihood ratio test rejects H_0 if

$$\lambda(X, Y) = \frac{\sup_{\underline{\theta} \in \Omega_0} L(\hat{\underline{\theta}} \mid X, Y)}{\sup_{\underline{\theta} \in \Omega} L(\hat{\underline{\theta}} \mid X, Y)}$$

$$= (1-r^2)^{n/2} < c$$

or equivalently, if

$$|r| > C_1$$

where C_1 is chosen that

$$\sup_{\underline{\theta} \in \Omega_0} P_{\underline{\theta}} \left[|r| > C_1 \right] = \alpha$$

The distribution of r under H_0 can be determined. It is known that

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

under null hypothesis has a students' t distribution with $n-2$ degrees of freedom. Using this fact the value of c_1 can be computed for a given α.

E3) There is another possible procedure of obtaining confidence intervals. This method has universal applicability and can be used for large or small samples. But this procedure usually yields confidence intervals that are much too large. The method uses the well known Chebychev's inequality

$$P\left\{|X - E(X)| < \varepsilon\sqrt{Var(X)}\right\} > 1 - \frac{1}{\varepsilon^2}, \text{ for } \varepsilon > 0$$

Let $\overline{X} = \frac{1}{n}\sum_{1}^{n} X_i$. We know that $E(\overline{X}) = p$ and $Var(\overline{X}) = p(1-p)/n$.

It follows that

$$P\left\{|\overline{X} - p| < \varepsilon\sqrt{p(1-p)/n}\right\} > 1 - \frac{1}{\varepsilon^2}$$

Since $p(1-p) \leq 1/4$, we have

$$P\left\{|\overline{X} - \varepsilon/2\sqrt{n} < p < \overline{X} + < \varepsilon/2\sqrt{n}\right\} > 1 - \frac{1}{\varepsilon^2}$$

One can now choose ε and n . Otherwise if n is fixed one can choose ε to get the desired level confidence interval.

APPENDIX

Table Values of the Standard Normal Distribution Function

$$\phi(z) = \int_{-x}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = P(Z \le z)$$

z	0	1	2	3	4	5	6	7	8	9
-3.0	0.0013	0.0010	0.0007	0.0005	0.0003	0.0002	0.0002	0.0001	0.0001	0.0000
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0026	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0084
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0119
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0238	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0300	0.0294
-1.7	0.0446	0.0436	0.0426	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0085	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0570	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.01711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2909	0.2061	0.2033	0.2266	0.1977	0.1949	0.1922	0.1894	0.1857
-0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2005	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2276
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3281	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.

* B.W. Lindgren, Statistical Theory, The Macmillan Company, 1960.

Table (Contd.)

$$\phi(z) = \int_{-x}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = P(Z \le z)$$

z	0	1	2	3	4	5	6	7	8	9
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	0.55910	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
4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
	9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
	9778	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9317
	9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
	.9861	0.9864	0.9868	0.9871	0.9874	0.9878	0.9881	0.9984	0.9887	0.9890
	9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
	0.9953	0.9955	0.9956	0.9956	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
	0.9871	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Table of χ^2 Distribution

If X has a χ^2 distribution with n degrees of freedom, this table gives the value of x such that $Pr[X \leq x] = p$.

n \ p	.005	.01	.025	.05	.10	.20	.25	.30	.40
1	.0000	.0002	.0010	.0039	.0158	.0642	.1015	.1484	.2750
2	.0100	.0201	.0506	.1026	.2107	.4463	.5754	.7133	1.022
3	.0717	.1148	.2158	.3518	.5844	1.005	1.213	1.424	1.869
4	.2070	.2971	.4844	.7107	1.064	1.649	1.923	2.195	2.753
5	.4117	.5543	.8312	1.145	1.610	2.343	2.675	3.000	3.655
6	.6757	.8721	1.237	1.635	2.204	3.070	3.455	3.828	4.570
7	.9893	1.239	1.690	2.167	2.833	3.822	4.255	4.671	5.493
8	1.344	1.647	2.180	2.732	3.490	4.594	5.071	5.527	6.423
9	1.735	2.088	2.700	3.325	4.168	5.380	5.899	6.393	7.357
10	2.156	2.588	3.247	3.940	4.865	6.179	6.737	7.267	8.295
11	2.603	3.053	3.816	4.575	5.578	6.989	7.584	8.148	9.237
12	3.074	3.571	4.404	5.226	6.304	7.807	8.438	9.034	10.18
13	3.565	4.107	5.009	5.892	7.042	8.634	9.299	9.926	11.13
14	4.075	4.660	5.629	6.571	7.790	9.467	10.17	10.82	12.08
15	4.601	5.229	6.262	7.261	8.547	10.31	11.04	11.72	13.03
16	5.142	5.812	6.908	7.962	9.312	11.15	11.91	12.62	13.98
17	5.697	6.408	7.564	8.672	10.09	12.00	12.79	13.53	14.94
18	6.265	7.015	8.231	9.390	10.86	12.86	13.68	14.43	15.89
19	6.844	7.633	8.907	10.12	11.65	13.72	14.56	15.35	16.85
20	7.434	8.260	9.591	10.85	12.44	14.58	15.45	16.27	17.81
21	8.034	8.897	10.28	11.59	13.24	15.44	16.34	17.18	18.77
22	8.643	9.546	10.98	12.34	14.04	16.31	17.24	18.10	19.73
23	9.260	10.20	11.69	13.09	14.85	17.19	18.14	19.02	20.69
24	9.886	10.86	12.40	13.85	15.66	18.06	19.04	19.94	21.65
25	10.52	11.52	13.12	14.61	16.47	18.94	19.94	20.87	22.62
30	13.79	14.95	16.79	18.49	20.60	23.36	24.48	25.51	27.44
40	20.71	22.16	24.43	26.51	29.05	32.34	33.66	34.87	36.16
50	27.99	29.71	32.36	34.76	37.69	41.45	42.94	44.31	46.86
60	35.53	37.48	40.48	43.19	46.46	50.64	52.29	53.81	56.62
70	43.27	45.44	48.76	51.74	55.33	59.90	61.70	63.35	66.40
80	51.17	53.54	57.15	60.39	64.28	69.21	71.14	72.92	76.19
90	59.20	61.75	65.65	69.13	73.29	78.56	80.62	82.51	85.99
100	67.33	70.06	74.22	77.93	82.86	87.95	90.13	92.13	95.81

Adapted with permission from *Biometrika Tables for Statisticians*, Vol. 1 3rd ed. Cambridge University Press, 1966, edited by E. S. Pearson and H.O. Hartley: and from "A new table of percentage point of the chi-square distribution." *Biometrika*, Vol. 51(1964), pp. 231 - 239, by H.L. Harter, Aerospace Research Laboratories.

Table of χ^2 Distribution (*Continued*)

.50	.60	.70	.75	.80	.90	.95	.975	.99	.995
.4549	7083	1.074	1.323	1.642	2.706	3.841	5.024	6.635	7.879
1.386	1.833	2.408	2.773	3.219	4.605	5.991	.7.378	9.210	10.60
2.366	2.946	3.665	4.108	4.642	6.251	7.815	9.348	11.34	12.84
3.357	4.045	4.878	5.385	5.989	7.779	9.488	11.14	13.28	14.86
4.351	5.132	6.064	6.626	7.289	9.236	11.07	12.83	15.09	16.75
5.348	6.211	7.231	7.841	8.558	10.64	12.59	14.45	16.81	18.55
6.346	7.283	8.383	9.037	9.803	12.02	14.07	16.01	18.48	20.28
7.344	8.351	9.524	10.22	11.03	13.36	15.51	17.53	20.09	21.95
8.343	9.414	10.66	11.39	12.24	14.68	16.92	19.02	21.67	23.59
9.342	10.47	11.78	12.55	13.44	15.99	18.31	20.48	23.21	25.19
10.34	11.53	12.90	13.70	14.63	17.27	19.68	2..92	24.72	26.76
11.34	12.58	14.01	14.85	15.81	18.55	21.03	23.34	26.22	28.30
12.34	13.64	15.12	15.98	16.98	19.81	22.36	24.74	27.69	29.82
13.34	14.69	16.22	17.12	18.15	21.06	23.68	26.12	29.14	31.32
14.34	15.73	17.32	18.25	19.31	22.31	25.00	27.49	30.58	32.80
15.34	16.78	18.42	19.37	20.47	23.54	26.30	28.85	32.00	34.27
16.34	17.82	19.51	20.49	21.61	24.77	27.59	30.19	33.41	35.72
17.34	18.87	20.60	21.60	22.76	25.99	28.87	3:.53	34.81	37.16
18.34	19.91	21.69	22.72	23.90	27.20	30.14	32.85	36.19	38.58
19.34	20.95	22.77	23.83	25.04	28.41	31.41	34.17	37.57	40.00
20.34	21.99	23.86	24.93	26.17	29.62	32.67	35.48	38.93	41.40
21.34	23.03	24.94	26.04	27.30	30.81	33.92	36.78	40.29	42.80
22.34	24.07	26.02	27.14	28.43	32.01	35.17	38.08	41.64	44.18
23.34	25.11	27.10	28.24	29.55	33.20	36.42	39.36	42.98	45.56
24.34	26.14	28.17	29.34	30.68	34.38	37.65	40.65	44.31	46.93
29.34	31.32	33.53	34.80	36.25	40.26	43.77	46.98	50.89	53.67
39.34	41.62	44.16	45.62	47.27	51.81	55.76	59.34	63.69	66.77
49.33	51.89	54.72	56.33	58.16	63.17	67.51	71.42	76.15	79.49
59.33	62.13	65.23	66.98	68.97	74.40	79.08	83.30	88.38	91.95
69.33	72.36	75.69	77.58	79.71	85,53	90.53	95.02	100.4	104.2
79.33	82.57	86.12	88.13	90.41	96.58	101.9	106.6	112.3	116.3
89.33	92.76	96.52	98.65	101.1	107.6	113.1	118.1	124.1	128.3
99.33	102.9	106.9	109.1	111.7	118.5	124.3	129.6	135.8	140.2

Table of the t distribution

If X has a t distribution with n degrees of freedom, the table gives the value of x such that $P(X \leq x) = p$.

n	p = 55	.60	.65	.70	.75	.80	.85	.90	.95	.975	.99	.995
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.474	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.77
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.75
40	.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704
60	.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660
120	.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326	2.5

This table is taken from Table III of Fisher & Yates : *Statistical Tables for Biological, Agricult and Medical Research*, published by Longman Group Ltd. London (previously published by and Boyd Ltd., Edinburgh) and by permission of the authors and publishers.

Table of the 0.95 Quantile of the *F* Distribution

If X has an F distribution with $n1$ and n degrees of freedom the table gives the value of v such that $P_r(N \leq x) = 0.975$.

n \ n_1	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	248.0	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	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	5.86	5.80	5.75	5.72	5.69	5.66	5.63
5	6.61	5.69	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	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	3.51	3.44	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	3.22	3.15	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	3.01	2.94	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	2.85	2.77	2.70	2.66	2.62	2.58	2.54
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.20	2.16	2.11	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.99	1.95	1.90	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.26	2.21	2.16	2.01	1.93	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	1.92	1.84	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.36	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.75	1.66	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.36	2.21	2.10	2.01	1.94	1.88	1.83	1.67	1.57	1.46	1.39	1.32	1.22	1.00

Adapted with permission from *Biometrika Tables for Statisticians, Vol.* 1, 3rd ed., Cambridge University Press, 1966, edited by E.S. Pearson and H.O. Hartley

If X has an F distribution with m and n degrees of freedom, the table gives the value of x such that $P_r(X \leq x) = 0.975$

n \ m	1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	984.9	993.1	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	6.52	5.46	5.27	5.17	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.68	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.26	3.20	3.14	3.08
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.59	2.52	2.246	2.40
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.29	2.22	2.16	2.09
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.13	2.07	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	1.94	1.82	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.83	1.71	1.57	1.48	1.39	1.27	1.00

...n permission from *Biometrika Tables for Statisticians, Vol.* 1, 3rd ed., Cambridge University Press, 1966, edited by ...n and H.O. Hartley.

ERRATA

MTE-11, Block 1)

Page No.	Line No.	Should be
15	7	.... similarly the value 5.3
15	8	....between 5.25 and 5.35. Thus, ... as 4.6—5.3 in
15	9	.... and ends at 5.35 .....
34	E 1) b)	show that $F_i = n - F'_{i+1}$ (for $i \geq k - 1$)
51	last line	$s^2 = \frac{\dots}{\dots} = 133490.18$
61	2 (from below)	deleted
62	6 (from below)	$\bar{x} = A + c\, v'_1$, ...
66	14 (from below)	$sk_4^2 = \frac{m_3^2}{m_2^3}$
73	2 (from below)	last column gives the .....
84	6 (from below)	and $Y_i = \sum_j f_{ij} y_j$
90	Fig. 6	The top most should be Group 1, and the lowest group should be Group 3

82

NOTES

NOTES