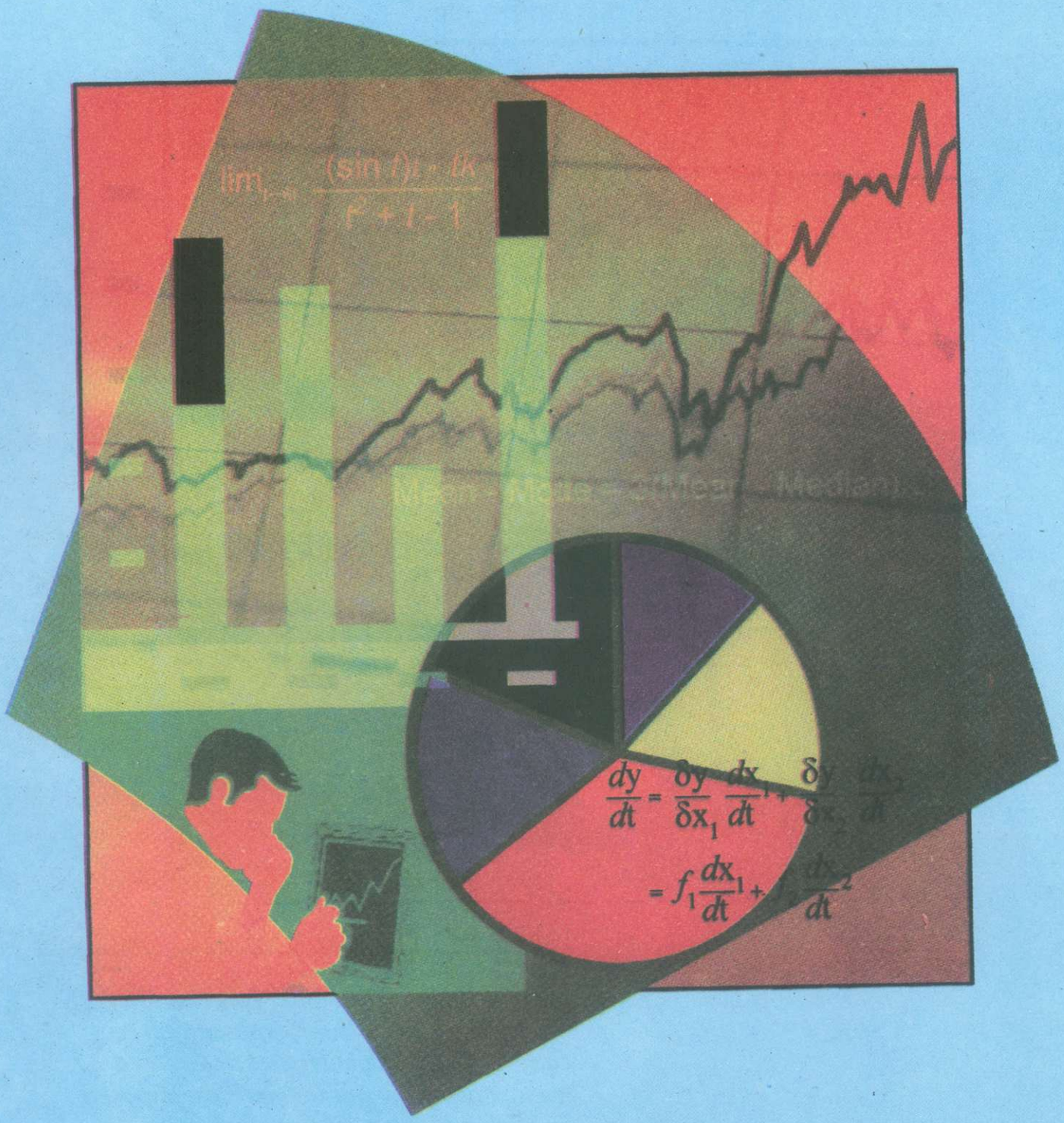
ignou
THE PEOPLE'S
UNIVERSITY

Indira Gandhi National Open University
School of Social Sciences

MEC-103
QUANTITATIVE METHODS



Statistical Methods-I

5

"शिक्षा मानव को बन्धनों से मुक्त करती है और आज के युग में तो यह लोकतंत्रा की भावना का आधार भी है। जन्म तथा अन्य कारणों से उत्पन्न जाति एवं वर्गगत विषमताओं को दूर करते हुए मनुष्य को इन सबसे ऊपर उठाती है।"

- इन्दिरा गाँधी

"Education is a liberating force, and in our age it is also a democratising force, cutting across the barriers of caste and class, smoothing out inequalities imposed by birth and other circumstances."

- Indira Gandhi

ignou
THE PEOPLE'S
UNIVERSITY

Indira Gandhi
National Open University
School of Social Sciences

MEC-103
Quantitative Methods

Block

5

STATISTICAL METHODS-I

UNIT 14
Descriptive Statistics 5

UNIT 15
Correlation and Regression Analysis 36

UNIT 16
Probability Distribution 71

UNIT 17
Probability Theory 95

MEC-103-5-15F-7K

Expert Committee

Prof. Bhaswar Moitra
Department of Economics
Jadavpur University
Kolkata

Dr. Naresh Kumar Sharma
University of Hyderabad
School of Economics, Hyderabad

Dr. Anirban Kar
Deptt. of Economics
Delhi School of Economics
University of Delhi

Dr. Indrani Roy Chowdhury
Economics Faculty, Jamia Millia Islamia
New Delhi

Prof. Sudhir Shah
Deptt. of Economics
Delhi School of Economics
University of Delhi

Dr. Manish Gupta
NIPFP,
Special Institutional Area, New Delhi

Prof. Gopinath Pradhan
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Prof. Narayan Prasad
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Prof. Kaustuva Barik
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Prof. B.S. Prakash
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Mr. Saugato Sen
School of Social Sciences
Indira Gandhi National Open University
New Delhi

Programme Coordinators

Prof. Gopinath Pradhan, IGNOU, New Delhi
Prof. Kaustuva Barik, IGNOU, New Delhi

Course Coordinator

Prof. Gopinath Pradhan

Course Editor

Prof. Gopinath Pradhan

Block Preparation Team

Unit	Writer
14	Sisir Debnath with updation by Gopinath Pradhan
15,16 & 17	Sisir Debnath

PRINT PRODUCTION TEAM

Mr. S. Burman	**Mr. Tilak Raj**	**Mr. Yashpal Sharma**
D.R.(P) MPDD-IGNOU,	A.R.(P) MPDD-IGNOU,	S.O.(P) MPDD-IGNOU,
Maidan Garhi, Delhi-110068	Maidan Garhi Delhi-110068	Maidan Garhi Delhi-110068

November 2018 (Reprint)

© Indira Gandhi National Open University, 2016

ISBN : 978-93-86375-16-2

All rights reserved. No Part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

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 or visit University's http:// www.ignou.ac.in.

Printed and Published on behalf of the Indira Gandhi National Open University, New Delhi by Registrar, MPDD. IGNOU

Printed at: A-One Offset Printers, 5/34, Kirti Nagar Indl. Area, New Delhi-110015

BLOCK 5 STATISTICAL METHODS-I

Important statistical tools often used in empirical and theoretical analyses are discussed in this block. Four broad themes around which we focus attention are related to data tabulation along with the preliminary descriptive statistics and summary measures, association between variables, probability theory and probability distribution. **Unit 14** contains discussion on basic sources of data and their tabular and graphical presentation, measures of location (such as mean, median and mode) and dispersion (standard deviation and variance). Simple analyses that can be undertaken from the observed association between two variables with the help of techniques like correlation and regression are presented in **Unit 15**. Basic elements of probability theory including additive and multiplicative theorems; conditional probability and independent events; Bayes' theorem and mathematical expectation are covered in **Unit 16**. The last unit (i.e., **Unit 17**) of the block deals with the concepts of mass and density, moment generating functions of probability along with important distributions such as binomial, poisson and normal.

BLOCK 5 STATISTICAL METHODS-I

Important statistical tools often used in empirical and theoretical analyses are discussed in this block. Four broad themes around which we focus attention are related to data tabulation along with the preliminary descriptive statistics and summary measures, association between variables, probability theory and probability distribution. **Unit 14** contains discussion on basic sources of data and their tabular and graphical presentation, measures of location (such as mean, median and mode) and dispersion (standard deviation and variance). Simple analyses that can be undertaken from the observed association between two variables with the help of techniques like correlation and regression are presented in **Unit 15**. Basic elements of probability theory including additive and multiplicative theorems; conditional probability and independent events; Bayes' theorem and mathematical expectation are covered in **Unit 16**. The last unit (i.e., **Unit 17**) of the block deals with the concepts of mass and density, moment generating functions of probability along with important distributions such as binomial, poisson and normal.

UNIT 14 DESCRIPTIVE STATISTICS

Structure

14.0 Objectives

14.1 Introduction

14.2 Data Presentation

- 14.2.1 Data: Types and Collection
- 14.2.2 Tabular Presentation
- 14.2.3 Charts and Diagrams for Ungrouped Data
- 14.2.4 Frequency Distribution
- 14.2.5 Histogram, Frequency Polygon and Ogives

14.3 Descriptive Statistics

- 14.3.1 Measures of Location
- 14.3.2 Measures of Dispersion
- 14.3.3 Measures of Skewness and Kurtosis

14.4 Let Us Sum Up

14.5 Key Words

14.6 Some Useful Books

14.7 Answer or Hints to Check Your Progress

14.8 Exercises

14.0 OBJECTIVES

After going through this unit, you will be able to

- tabulate data from primary and secondary sources; and
- analyse data using some of frequently used statistical measures.

14.1 INTRODUCTION

Statistics in the broadest sense refers to a range of techniques and procedures for analyzing, interpreting displaying and decision making based on a set of data. These are a set of facts and figures collected by an individual or an authority on the concerned topic. These data collected are often a huge mass of haphazard numerical figures and you need to present them in a comprehensive and systematic fashion amenable to analysis. For that purpose, we are introduced to data presentation and preliminary data analysis in the following discussion.

14.2 DATA PRESENTATION

In this section, we will discuss some useful ways of compiling data. Before that, we will introduce some basic concepts of types of data and methods of data collection.

14.2.1 Data Types and Collection

Data are systematic record of values taken by a variable or a number of variables on a particular point of time or over different points of time.

Data collected on a single point of time over different sections (may be classified on demographic, geographic or other considerations) are called

cross-section data. Whereas data collected over a period of time are called time series data.

Data may be quantitative or qualititative in nature. For example, heights of 50 students of Delhi University are quantitative whereas religion of theirs is qualititative in nature. Data of quantitative nature are technically called variables whereas data of qualitative nature are called attributes. Again, variables may be discrete as well as continuous. If a variable can take any value within its range, then it is called a continuous variable otherwise it is called a discrete variable. Heights of students of Delhi University are a continuous variable whereas number of students under different Universities of India is discrete variable.

Depending on the type of collection, data may be of two types, namely, Primary data and Secondary data. Primary data are those which are collected for a specific purpose directly from the field of enquiry and hence they are original in nature. On the other hand, data collected by someone but used by another or collected for one purpose and used for another are called secondary data. Following are few examples of primary and secondary data.

Primary data:

1) "Reserve Bank of India Bulletin", published monthly by Reserve Bank of India.

2) "Indian Textile Bulletin", issued monthly by Textile Commissioner, Mumbai.

Secondary data:

1) "Monthly Abstract of Statistics".

By whatever means data are collected or classified, they need to be presented so as to reveal the hidden facts or to ease the process of comprehension of the field of enquiry. Generally, data are presented by the means of

i) Tables and

ii) Charts and Diagrams.

14.2.2 Tabular Presentation of Data

Tabulation of data may be defined as the logical and systematic organisation of statistical data in rows and columns, designed to simplify the presentation and to facilitate quick comparison. In tabular presentation errors and emissions could be readily detected. Another advantage of tabular presentation is avoidance of repetition of explanatory terms and phrases. A table constructed for presenting the data has the following parts:

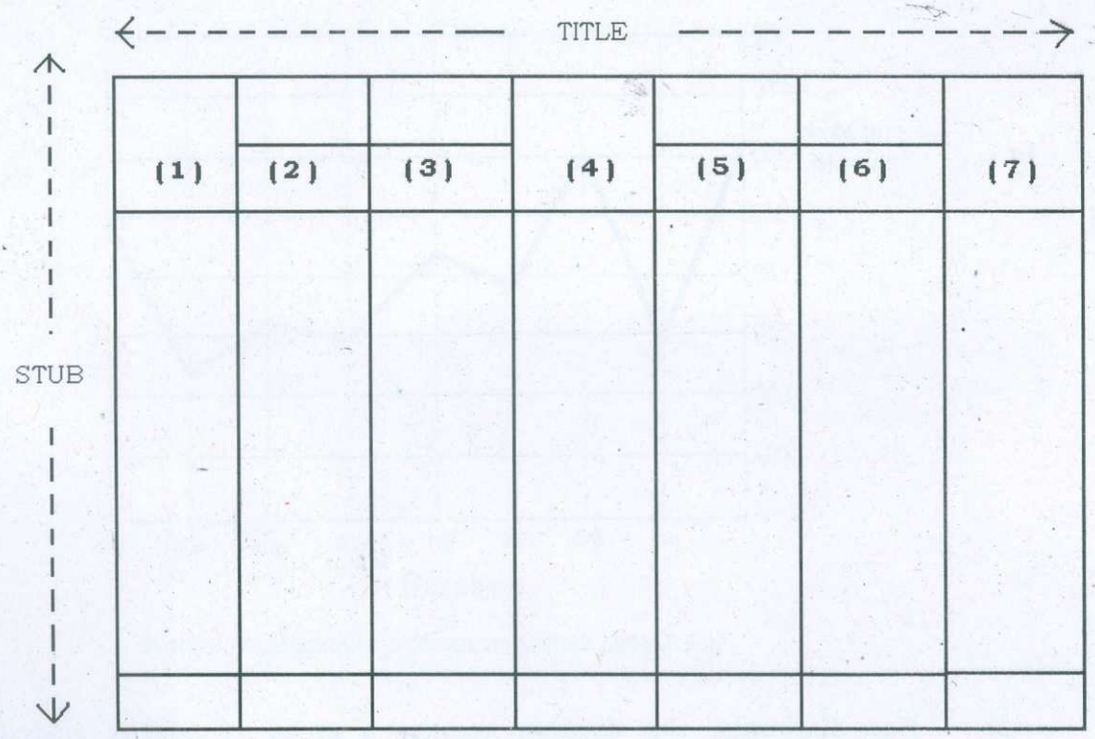
1) **Title:** This is brief description of the contents and is shown on the top of the table.

2) **Stub:** The extreme left part of a table is called Stub. Here the descriptions of the rows are shown.

3) **Caption and Box Head:** The upper part of the table, which shows the description of columns and sub columns, is called Caption. The row of the upper part, including caption, units of measurement and column number, if any, is called box-head.

4) **Body:** This part of the table shows the figures.

5) **Footnote:** In this part we show the source of data and explanations, if any.



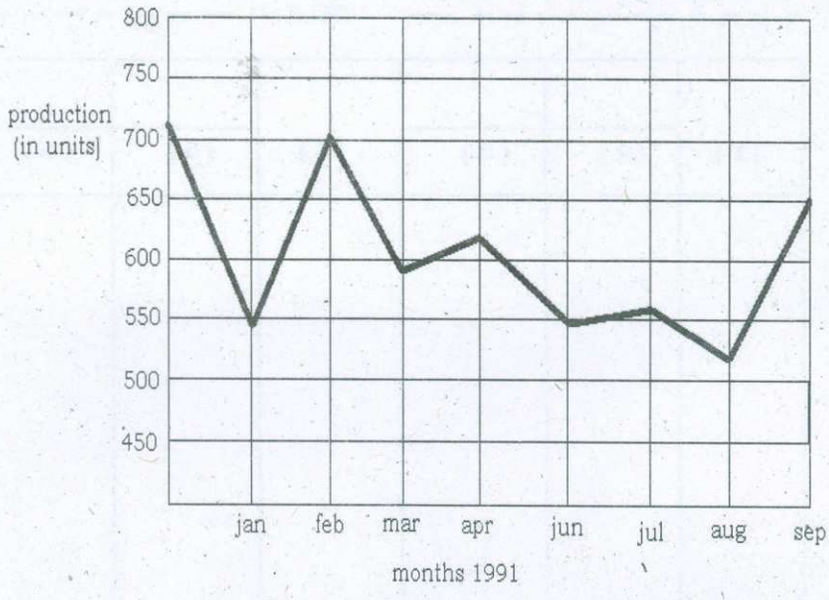
14.2.3 Charts and Diagrams for Ungrouped Data

Charts and diagrams are useful devices for the data presentation. Diagrams are appealing to the eyes and they are helpful in assimilating data readily and quickly. A chart, on the other hand, can clarify complex problems and reveal the hidden facts. But charts or diagrams unlike tables do not show details of data and require much time to construct. Note that the words 'charts' and 'diagrams' are used almost in the same sense. The common types of charts and diagrams are,

1) Line diagrams.
2) Bar diagrams.
3) Pie diagrams.
4) Pictogram.

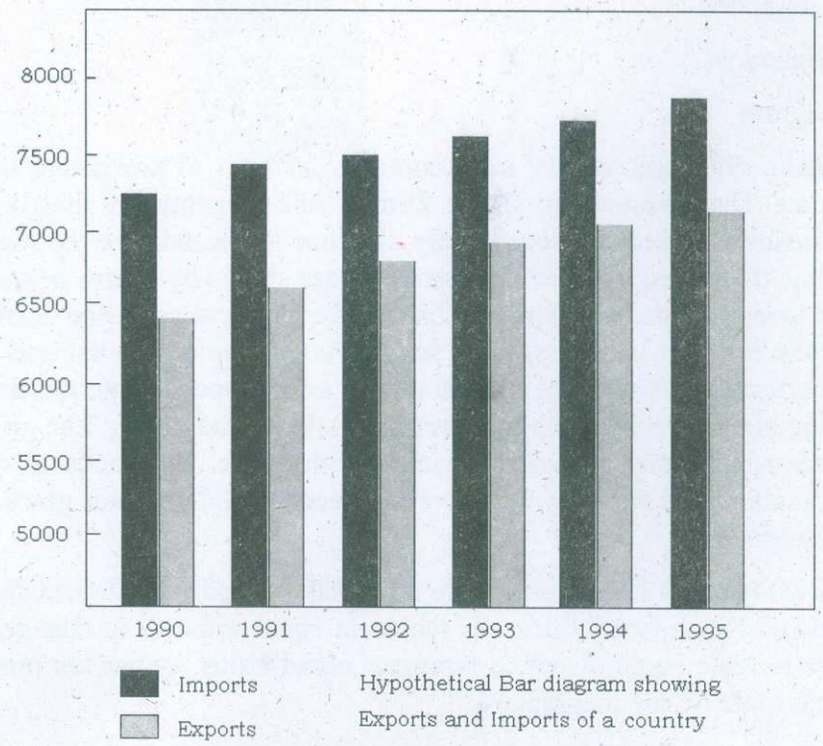
1) **Line diagrams** are the most common methods of presenting statistical data. Data presentation in the form of line diagrams are mostly used in business and commerce. Mostly, the time series data are represented by line diagrams. In a line diagram, data are shown by means of a curve or a straight line. The straight line or the curve reveals the relationship between two variables. Two straight lines, one horizontal and another vertical (known as the X axis and Y axis, respectively), are drawn on the graph paper, which intersect at a point called origin. The given data are represented as points on the graph paper. The locus of all such points joined either by curves or by pieces of straight lines gives the line diagram.

   Two types of line diagrams are used, natural scale and ratio scale. In the natural scale equal distances represent equal amounts of change. But in ratio scale equal distances represent equal ratios. Below we provide an example of line diagram.



Line diagram showing production of a firm against months of 1991

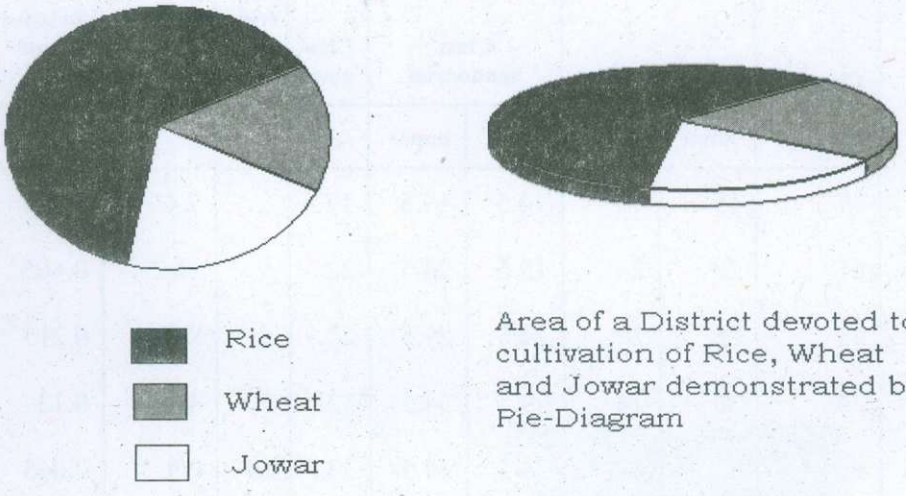
2) **Bar diagrams:** Bar diagram consists a group of equally spaced rectangular bars, one for each category (or class) of given statistical data. The rectangular bars are differentiated by different shades or colors. The bars starting from a common baseline must be of equal width and their length represents the values of statistical data. Bar diagrams may be of two types: vertical and horizontal. For each of these types, we have again grouped bar diagram, subdivided bar diagram, paired bar diagram etc. Grouped bar diagrams are used to show the comparison of two or more sets of related statistical data, while subdivided or component bar diagrams are used for comparing the sizes of the different component parts among themselves. The paired bar diagram consists of several pairs of horizontal bars. Following figure shows a paired bar diagram.



Hypothetical Bar diagram showing Exports and Imports of a country

3) **Pie diagrams:** A pie diagram is a circle whose area is divided proportionately among the different components by straight lines drawn from the center to the circumference. When statistical data are given for a number of categories and we are interested in their comparison in a manner that reveals the contribution of each category to the total, pie diagrams are very useful in effectively displaying the data.

In a pie diagram, it is necessary to express the value of each category as a percentage of the total. Since the circle represents the total and to represent each category in that circle, we have to multiply the percentage of each category by 3.6 degrees, so that sum of each category becomes 360 degrees. The diagram can be drawn with the help of a compass and a protractor. The following figure shows a hypothetical Pie diagram.



Area of a District devoted to cultivation of Rice, Wheat and Jowar demonstrated by a Pie-Diagram

4) **Pictogram:** This type of data presentation consists of rows of pictures or symbols of equal size. Each picture or symbol represents a definite numerical value. Pictograms help to present data to illiterate people or to children.

14.2.4 Frequency Distribution

Frequency of a variable is the number of times it occurs in given data. Suppose we have data on the daily number of accidents in a city for 30 days. If 5 accidents have occurred 6 times in these 30 days, then frequency of 5 accidents daily is 6. Thus, if we have a large mass of data we can compress these by writing the frequency of each variable corresponding to the values or the range of values taken by these. Let us suppose that the variable x takes the values as $x_1, x_2 \ldots x_n$. Then frequency of x_i is generally denoted by f_i. There are two types of frequency distribution, namely, simple frequency distribution and grouped frequency distribution. Simple frequency distribution shows the values of the variable individually whereas the grouped frequency distribution shows the values of the variable in groups or intervals. Following two tables will elucidate on the different types of frequency distributions.

Table 1 shows simple frequency distribution of number of problems solved by a student daily during a month.

Table 1: Number of Problems Solved by a Student Daily (during a month)

Number of problems solved	Frequency
3	5
4	6
5	4
6	10
7	5
Total	**30**

Table 2 shows grouped frequency distribution of a hypothetical data.

Table 2: Grouped Frequency Distribution of Age in a Locality

Class Interval	Class frequency	Class limit		Class boundaries		Class mark	Width of Class	Frequency Density	Relative Frequency
		lower	upper	lower	upper				
15 –19	37	15	19	14.5	19.5	17	5	7.4	0.185
20 – 24	81	20	24	19.5	24.5	22	5	16.2	0.405
25 – 29	43	25	29	24.5	29.5	27	5	8.6	0.215
30 – 34	24	30	34	29.5	34.5	32	5	4.8	0.12
35 – 44	9	35	44	34.5	44.5	39.5	10	0.9	0.045
45 – 59	6	45	59	44.5	59.5	52	15	0.4	0.03

In the example of grouped frequency distribution, we have shown few useful terms associated with it. They are class intervals or class, class frequency, cumulative frequency (greater than and less than type), class limits (upper and lower), class boundaries (upper and lower), mid point of class interval, width of a class, relative frequency and lastly frequency density. We will formally define these terms.

Class: When a large number of observations varying in a wide range are available, they are usually classified into several groups according to the size of the values. Each of these groups defined by an interval is called class interval or simply class.

Class Frequency: The number of observation falling under each class is called its class frequency or simply frequency.

Class Limits: The two numbers used to specify the limits of a class interval for tallying the original observations are called the class limits.

Class Boundaries: The extreme values (observations) of a variable, which could ever be included in a class interval, are called class boundaries.

Mid-Point of Class Interval: The value exactly at the middle of a class interval is called class mark or mid-value. It is used as the representative value of the class interval. Thus, Mid-point of Class interval = (Lower class boundary +Upper class boundary)/2.

Width of a Class: Width of class is defined as the difference between the upper and lower class boundaries. Thus, Width of a Class = (upper class boundary – lower class boundary).

Relative Frequency: The relative frequency of a class is the share of that class in total frequency. Thus, Relative Frequency = (Class frequency / Total frequency).

Frequency Density: Frequency density of a class is its frequency per unit width. Thus, Frequency density = (Class frequency / Width of the class).

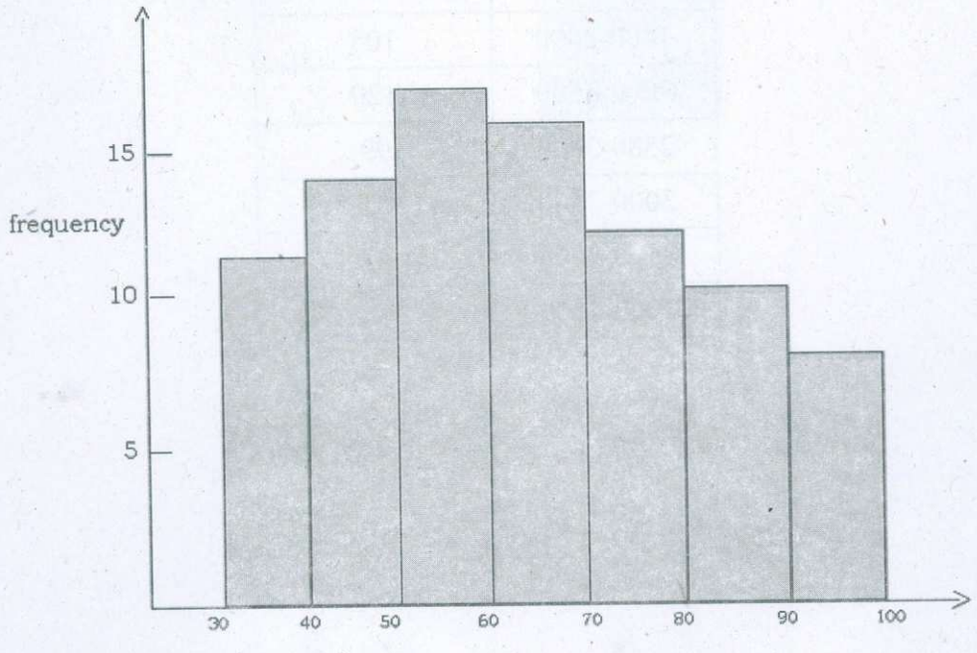
Cumulative Frequency: Cumulative frequency corresponding to a specified value of a variable or a class (in case of grouped frequency distribution) is the number of observations smaller (or greater) than that value or class. The number of observation up to a given value (or class) is called less-than type cumulative frequency distribution, whereas the number of observations greater than a value (or class) is called more-than type cumulative frequency distribution.

14.2.5 Histogram, Frequency Polygon and Ogives

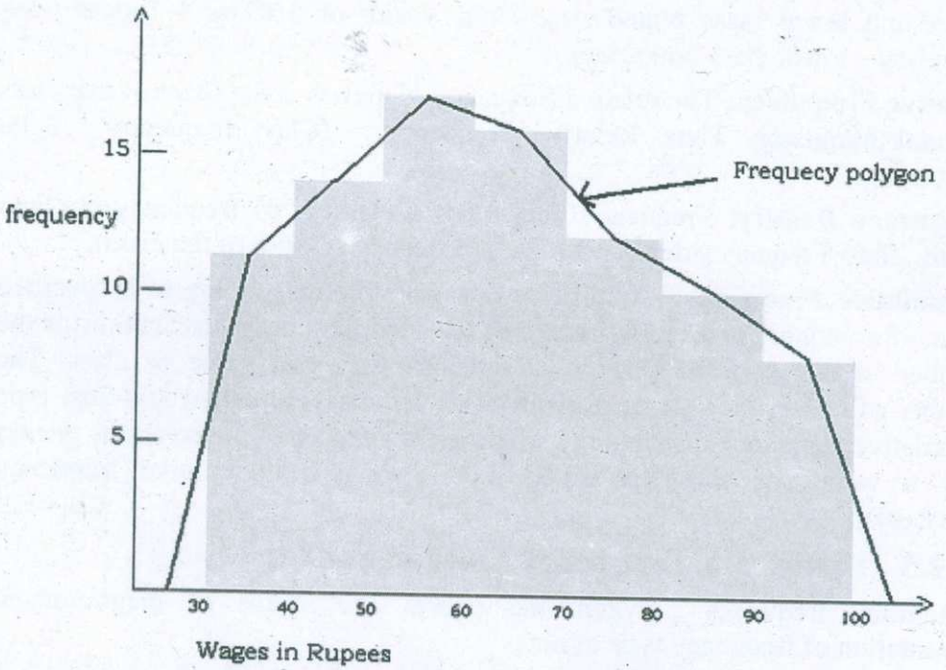
Histogram, frequency polygon and ogives are means of diagrammatic presentation of frequency type of data.

1) **Histogram** is the most common form diagrammatic presentation of grouped frequency data. It is a set of adjacent rectangles on a common base line. The base of each rectangle measures the class width whereas the height measures the frequency density.
2) **Frequency Polygon** of a frequency distribution could be achieved by joining the midpoints of the tops of the consecutive rectangles. The two end points of a frequency polygon are joined to the base line at the mid values of the empty classes at the end of the frequency distribution.
3) **Ogives** are nothing but the graphical representation of the cumulative distribution. Plotting the cumulative frequencies against the mid-values of classes and joining them, we obtain ogives.

Following are the examples of histogram, frequency polygon and ogives.



Histogram showing distribution casual labor in a locality



Frequency polygon showing distribution casual labor in a locality

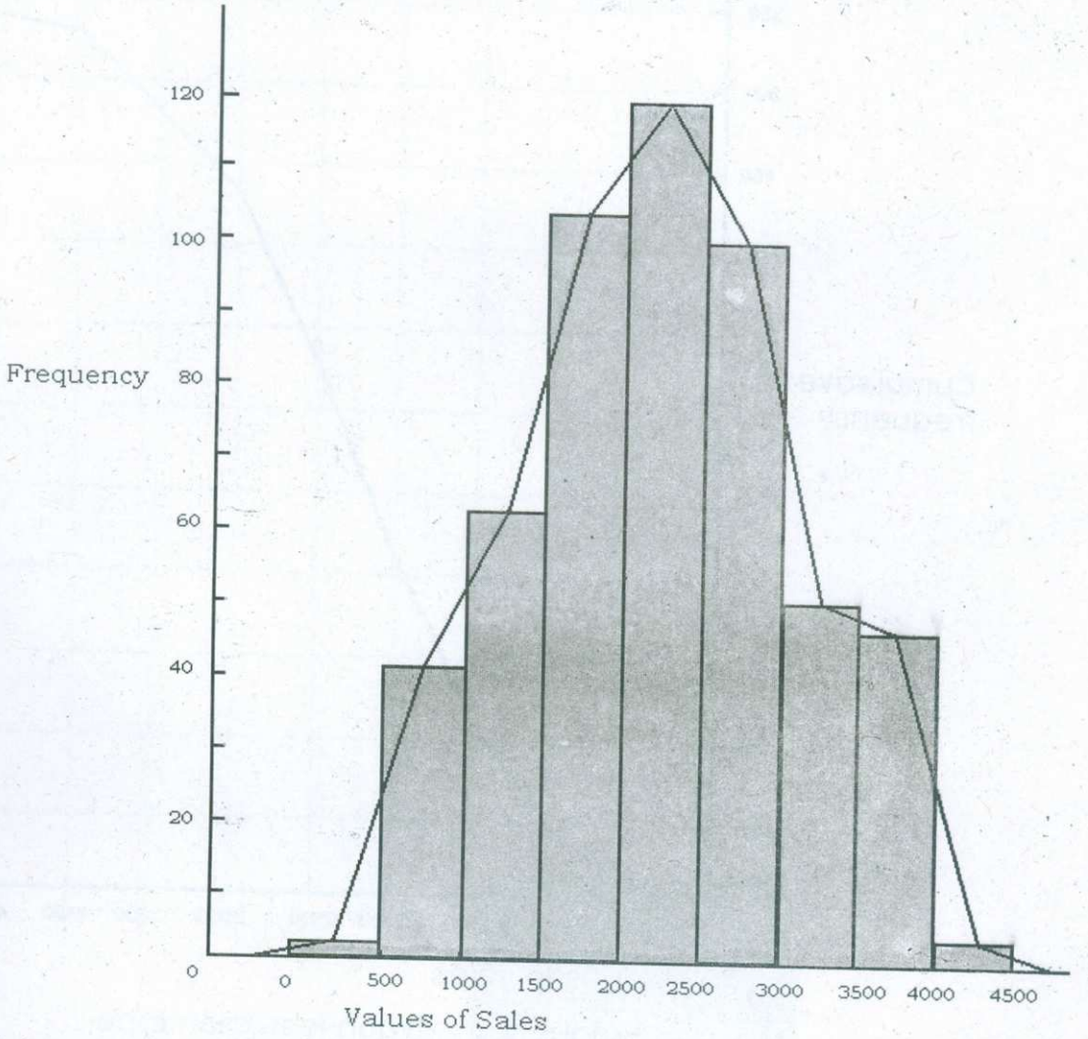
Example: Following data were obtained from a survey on the value of annual sales of 534 firms. Draw the histogram and the frequency polygon and ogive from the data.

Table 3 : Value of Annual Sales of 534 firms

Value of sales	Number of firms
0-500	3
500-1000	42
1000-1500	63
1500-2000	105
2000-2500	120
2500-3000	99
3000-3500	51
3500-4000	47
4000-4500	4

Solution:

It is relatively easier to draw histogram and frequency polygon.

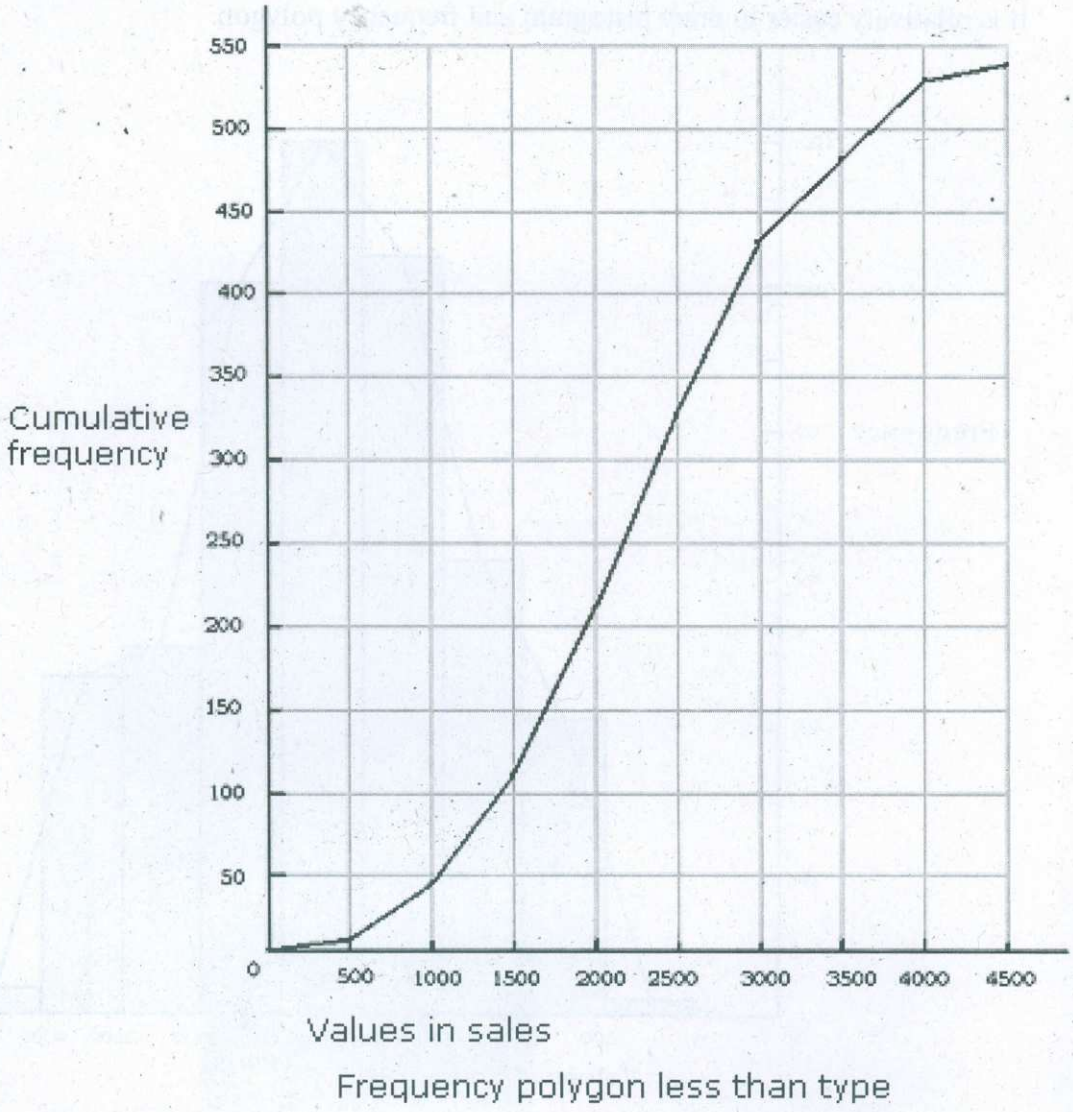


In order to draw the frequency polygon we have to construct the cumulative frequency distribution from the above data. It is done in the next table.

Table 4 : Cumulative Frequency of Annual Sales

Class Boundary	Cumulative Frequency
0	0
500	3
1000	45
1500	108
2000	213
2500	333
3000	432
3500	483
4000	530
4500	534

We plot the above to get the following given.



Frequency polygon less than type

Check Your Progress 1

1) Explain the advantages of tabular presentation of data over textual presentation.

2) Prepare a blank table showing the average height of males and females classified into two age groups of eighteen years and over, and under eighteen years in seven districts on the years 2004 and 2005.

3) Represent the following data by line diagram and bar diagram.

Year	Value of Exports (Rs. crore)	Value of Imports (Rs. crore)
1937 - 1938	301	243
1938 - 1939	295	226
1939 - 1940	309	230
1940 - 1941	260	184
1941 - 1942	276	168
1942 - 1943	184	85
1943 - 1944	158	89
1944 - 1945	156	160
1945 - 1946	182	177

4) Draw a pie chart for the following data on cotton exports

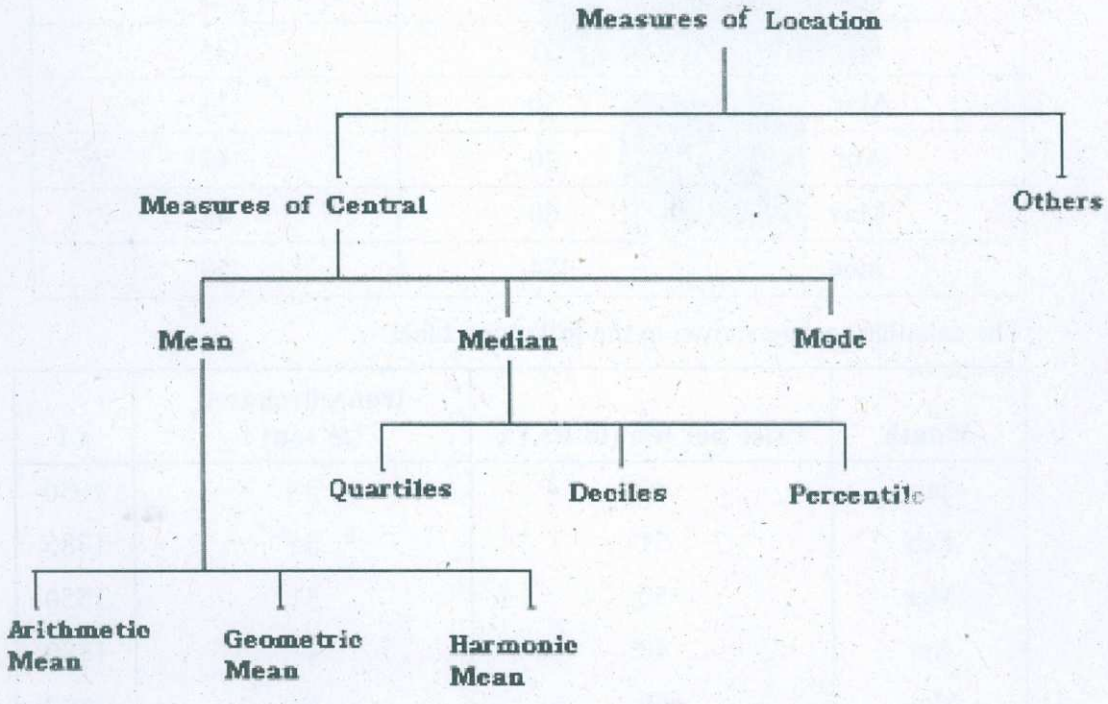
Country	Exports of Cotton (in bales)
U.S.A	6367
India	2999
Egypt	1688
Brazil	650
Argentina	202
Total	11906

14.3 DESCRIPTIVE STATISTICS

The collection, organisation and graphic presentation of numerical data help to describe and present these into a form suitable for deriving logical conclusions. Analysis of data is another way to simplify quantitative data by extracting relevant information from which summarized and comprehensive numerical measures can be calculated. Most important measures for this purpose are measures of location, dispersion and symmetry and skewness. In this section we will discuss these measures in the order just stated.

14.3.1 Measures of Location

A single value can be derived for a set of data to describe the elements contained in it. Such a value is called a measure of location. Again, the central tendency is the property of data by virtue of which they tend to cluster around some central part of the distribution. Mean, median and mode are the measures of central tendency. There are other measures of location, namely, quartiles, deciles and percentiles. The following figures show is a summary of measures of location



Measures of Central Location

Mean

Arithmetic mean

Arithmetic mean of a set of realisations of a variable is defined as their sum divided by the number of observations. It is usually denoted by $\bar{x}$ (read as x bar) where x denotes the variable. Depending on whether the data are grouped or ungrouped arithmetic mean may be of two types. First, simple arithmetic mean for ungrouped data and second, weighted arithmetic mean for grouped (frequency type) data. If the realisations of the variable x are $x_1, x_2 \ldots x_n$ than,

Simple Arithmetic Mean $(\bar{x}) = (x_1 + x_2 + \ldots\ldots + x_n) / n$

$= \sum_{i=1}^{n} \frac{x_i}{n}$ [$\sum$ is the summation operator which sums over different values taken by a variable] . If the variable x takes the values $x_1, x_2 \ldots x_n$ with frequencies $f_1, f_2 \ldots f_n$ then

Weighted arithmetic mean $(\bar{x}) = x_1 f_1 + x_2 f_2 + \ldots\ldots + x_n f_n / \sum_{i=1}^{n} f_i = \frac{\sum_{i=1}^{n} x_i fi}{\sum_{i=1}^{n} f_i}$.

Example: *Given the following data calculate the simple and weighted arithmetic average price per ton of iron purchased by an industry for six months.*

Month	Price per ton (in Rs.)	Iron purchased (in ton)
Jan.	42	25
Feb.	51	35
Mar.	50	31
Apr.	40	47
May	60	48
June	54	50

The calculations are shown in the following table

Month	Price per ton (in Rs.) x	Iron purchased (in ton) f	x.f
jan	42	25	1050
Feb	51	35	1785
Mar	50	31	1550
Apr	40	47	1880
May	60	48	2880
Jun	54	50	2700
Total	297	236	11845

Simple Arithmetic Mean = $\sum_{i=1}^{n} x_i$ / n = 297 / 6 = 49.5

Weighted Arithmetic Mean = $\sum_{i=1}^{n} x_i f_i / \sum_{i=1}^{n} f_i$ = 11845 / 236 = 50.19

Given two groups of observations, n_1 and n_2, and $\bar{x}$ and $\bar{x}_2$ being the number of observations and arithmetic mean of two groups respectively, we can calculate the composite mean using the following formula:

Composite Mean ($\bar{x}$) = ($n_1 . \bar{x}_1 + n_2 . \bar{x}_{2)} / n_1 + n_2$

Geometric mean

Geometric mean of a set of observations is nth root of their product, where n is the number of observation. In case of non frequency type data, simple geometric mean

= $\sqrt[n]{}$ ($x_1 \times x_2 \times x_3 \times x_4 \times \ldots\ldots \times x_n$) and

in case of frequency type data weighted geometric mean

= $\sum_i f_i \sqrt{}$($x_1 f_1 \times x_2\ f_2 \times x_3\ f_3 \times x_4\ f_4 \times \ldots\ldots \times x_n f_n$).

Geometric mean is more difficult to calculate than arithmetic mean. However, since it is less affected by the presence of extreme values, it is used to calculate index numbers.

Example: *Apply the geometric mean to find the general index from the following group of indices by assigning the given weights.*

Group	X
A	118
B	120
C	97
D	107
E	111
F	93
	Total

Weighted geometric mean (g) = $\sum_{i=1}^{n} f_i \sqrt{}$($x_1 f_1 \times x_2\ f_2 \times x_3\ f_3 \times x_4\ f_4 \times \ldots\ldots \times x_n f_n$).

Taking logarithms we get log g = $1/\sum_{i=1}^{n} f_i \times \log x_i$

Thus, logarithm of weighted (simple) geometric mean is equal to the weighted (simple) A.M. of the logarithm of the observations.

X	f	log x	f × log x
118	4	2.071882	8.287528
120	1	2.079181	2.079181
97	2	1.986772	3.973543
107	6	2.029384	12.1763

111	5	2.045323	10.22661
93	2	1.968483	3.936966
Total	20		40.68014

Log g = $1/\sum_{i=1}^{n} f_i \times \log x_i = 40.68/20 = 2.03$.

Therefore,

g = antilog 2.03 = 108.1.

Harmonic mean

It is the reciprocal of the arithmetic mean and computed with the reciprocal of the observations. For data without frequency,

$$\text{simple harmonic mean} = \frac{n}{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)}.$$

In case of data with frequency,

$$\text{harmonic mean} = \left(\frac{\sum_{i=n}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}\right).$$

Example: *A person bought 6 rupees worth of mango from five markets at 15, 20, 25, 30 and 35 paise per mango. What is the average price of a mango?*

Average price is the H.M. of *15, 20, 25, 30 and 35.*

$$\text{Average price} = \frac{5}{\frac{1}{15}+\frac{1}{20}+\frac{1}{25}+\frac{1}{30}+\frac{1}{35}} = 1500/63 = 24\text{p.}$$

Harmonic mean has limited use. It gives the largest weight to the smallest observation and the smallest weight to the largest observation. Hence, when there are few extreme values present in the data, harmonic mean is preferred to any other measures of central tendency. It may be useful to note that harmonic mean is useful in calculating averages involving time, rate and price.

For a given set of observations the following inequality holds:

A.M. $\geq$ G.M. $\geq$ H.M.

Suppose there are only two observations x_1 and x_2,

$$\left(\sqrt{x_1} - \sqrt{x_2}\right)^2 \geq 0$$

or, $x_1 + x_2 - 2\sqrt{x_1}\sqrt{x_2} \geq 0$

or, $x_1 + x_2 \geq 2\sqrt{x_1}\sqrt{x_2}$

or, $(x_1 + x_2)/2 \geq \sqrt{x_1}\sqrt{x_2}$

or, A.M. $\geq$ G.M.

Similarly,

$$\left(\sqrt{\frac{1}{x_1}}-\sqrt{\frac{1}{x_2}}\right)^2 \geq 0$$

$$\text{or, } \frac{1}{x_1}+\frac{1}{x_2}-2\sqrt{\frac{1}{x_1}}\sqrt{\frac{1}{x_2}} \geq 0$$

$$\text{or, } \frac{1}{x_1}+\frac{1}{x_2} \geq 2\sqrt{\frac{1}{x_1}}\sqrt{\frac{1}{x_2}}$$

$$\text{or, } (\frac{1}{x_1}+\frac{1}{x_2})/2 \geq \sqrt{\frac{1}{x_1}}\sqrt{\frac{1}{x_2}}$$

$$\text{or, } \sqrt{x_1 x_2} \geq \frac{2}{\left(\frac{1}{x_1}+\frac{1}{x_2}\right)}$$

or, G.M. ≥ A.M.

Thus, we can prove A.M. ≥ G.M. ≥ H.M. for 2 observations. This result holds for any number of observations.

Median:

Median of a set of observation is the middle most value when the observations are arranged in order of magnitude. The number of observations smaller than median is the same as the number of observations greater than it. Thus, median divides the observations into two equal parts and in a certain sense it is the true measure of central tendency, being the value of the most central observation. It is independent of the presence of extreme values and can be calculated from frequency distributions with open-ended classes. Note that in presence of open-ended process calculation of mean is not possible.

Calculation of Median

a) For ungrouped data, the observations have to be arranged in order of magnitude to calculate median. If the number of observations is odd, the value all the middle most observation is the median. However, if the number is even, the arithmetic mean of the two middle most values is taken as median.

b) For simple frequency distribution, to calculate median, we have to calculate the less than type cumulative frequency distribution. If the total frequency is N, the value of the variable corresponding to the cumulative frequency $\frac{N+1}{2}$ gives the median.

c) Median of a grouped frequency distribution is that value of the variable which corresponds to the cumulative frequency N/2. Median can be calculated using either the formula or graph. Both these methods are given below:

1) To use the formula for median we have to calculate the cumulative frequency for each class. The class in which the cumulative frequency N/2 lies is called the median class. To compute median we apply the following formula:

Median = $l_1 + (N/2 - F / f_m) \times c$

where, l_1 : lower boundary of the median class

N : total frequency

F : cumulative frequency below l_1

F_m : frequency of the median class

c : difference between upper and lower class limits of the median intervals.

2) An appropriate value of median can be calculated graphically from ogives or cumulative frequency polygon. We have to draw a horizontal line from the point N/2 on the vertical axis, which shows the cumulative frequency, until it meets the ogives (either less than type or greater than type). From this point of intersection, a perpendicular is dropped on the horizontal axis. The position of the foot of the perpendicular is read from the horizontal scale showing the values of the variable. The advantage of median is that it is easy to understand and calculate. It could be calculated even if all the observations are not known. Median could also be calculated from grouped frequency distributions with classes of unequal width. But there are many disadvantages also. For the calculation of median data must be arranged. Unlike mean, median cannot be treated algebraically. In median, it is not possible to give higher weights to smaller values and smaller weights to higher values. Calculation of median from grouped frequency distribution assumes that the observations in the median class are uniform which may not be true always.

Example: *Find median and median class for the following data:*

15 – 25	25 - 35	35 - 45	45 - 55	55 – 65	65 - 75
4	11	19	14	0	2

Solution :

Median

class Boundary	Cumulative Frequency
15	0
25	4
35	15
45	34
55	48
65	48
75	50

N / 2 = 25

Since N/2 lies between the cumulative frequencies 15 and 34 , therefore median must lie in the interval between 35 and 45. Now applying simple interpolation

$$\frac{\text{median - 35}}{\text{45 - 35}} = \frac{\text{25 - 15}}{\text{34 -15}}$$

The above equation gives median = 40.26

Alternatively, we can use the following formula

Median = $l_1 + (N/2 - F / f_m) \times c$

class Boundary	Frequency	Cumulative Frequency
15 - 25	4	4
25 - 35	11	15 ← F
35 - 45	**19** ←	**34**
45 - 55	14	48
55 - 65	0	48
65 - 75	2	50

frequency of the median class or f_m

Using the above formula we will get the same result.

Mode:

Mode of a given set of observation is that value of the variable which occurs with the maximum frequency. Concept of mode is generally used in business as it is most likely to occur. Meteorological forecasts are based on mode.

From a simple series, mode can be calculated by locating that value which occurs maximum number of times.

From a simple frequency distribution mode can be determined by inspection only. It is that value of the variable which corresponds to the largest frequency.

From the grouped frequency distribution mode can be determined. It is very difficult to find the mode accurately. However, if all classes are of equal width, mode is usually calculated using the following formula:

$\text{Mode} = l_1 + \{d_1 / (d_1 + d_2)\} \times c$

where, l_1: lower boundary of the modal class (i.e., the class with the highest frequency)

d_1: difference of the largest frequency and the frequency of the class just preceding the modal class

d_2: difference of the largest frequency and the frequency of the class just following the modal class

c: common width of classes.

Example: *The number of telephone calls received in 245 successive one minute intervals at an exchange are shown in the following frequency distribution. Evaluate the mode.*

No. of calls	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Frequency	*14*	*21*	*25*	*43*	*51*	*40*	*39*	*12*

Mode is the value of the variable corresponding to the highest frequency, which is 51 in the problem. Therefore, mode is 4.

If however the frequency distribution has classes of unequal width the above formula cannot be applied. In that case, an approximate value of mode is obtained by the following relation between mean, median and mode.

Mean – Mode = 3 (Mean – Median), when mean and median are known.

There are many advantages of mode. From a simple frequency distribution mode can be calculated only by inspection. It is unaffected by the presence of extreme values of the observations and can be calculated from frequency distribution with open-ended classes. The disadvantages associated with it cannot also be overlooked, however. Mode has no significance unless large number of observations is available. When all values of the variable occur with equal frequency there is no mode. On the other hand, if two or more values have the same maximum frequency, then there is more than one mode. Unlike mean, mode cannot be treated algebraically.

For uni-modal distributions the following approximate relation holds:

Mean – Mode = 3 (Mean – Median)

Other Measures of Location

Just as median divides the total number of observations into two equal parts, there are other measures which divide the observations into fixed number of parts, say, 4 or 10 or 100. These are collectively known as partition values or quartiles. Some of them are,

1) quartiles 2) deciles and 3) percentiles.

Median which falls into this group has already been discussed. Quartiles are such values which divide the total observations into four equal parts. To divide a set of observations into four equal parts three dividers are needed. These are first quartile, second quartile and third quartile. The number of observations smaller than Q_1 is the same as the number of observations lying between Q_1 and Q_2, are between Q_2 and Q_3 or larger then Q_3. One quarter of the observations is smaller then Q_1, two quarter of the observations are smaller then Q_2 and three quarter of the observations are smaller then Q_3. This implies Q_1, Q_2, Q_3 are values of the variable when the less than type cumulative frequencies is N/4, N/2 and 3N/4 respectively. Clearly, $Q_1 < Q_2 < Q_3$; Q_2 stands for median (as half of the observations are greater than the median and rest half are smaller than it. In other words, median divides the observations into two equal parts)

Similarly, deciles divide the observations into ten equal parts and percentiles divide observations into 100 equal parts.

Check Your Progress 2

1) The number of telephone calls received in 245 successive one minute intervals at an exchange are shown below:

No. of calls	0	1	2	3	4	5	6	7
Frequency	14	21	25	43	51	40	39	12

Calculate the mean, median and median.

2) If A.M. = 26.8, Median = 27.9 find the value of mode.

3) Give examples of situations where mode is the appropriate measure of central tendency.

4) Explain the advantages and disadvantages of using mode as a measure of central location.

14.3.2 Measures of Dispersion

Measures of central location alone cannot adequately represent or summarize the statistical data because they do not provide any information concerning the spread of actual observations. Two sets of data having the same measure of central tendency do not necessarily exhibit the same kind of dispersion. The word dispersion is used to denote degree of heterogeneity in the data. It is an important characteristic indicating the extent to which the observations may vary among themselves. As an example of this, consider the following two series.

Series A	30	33	35	40	37	Total = **175**
Series B	2	21	31	58	63	Total = **175**

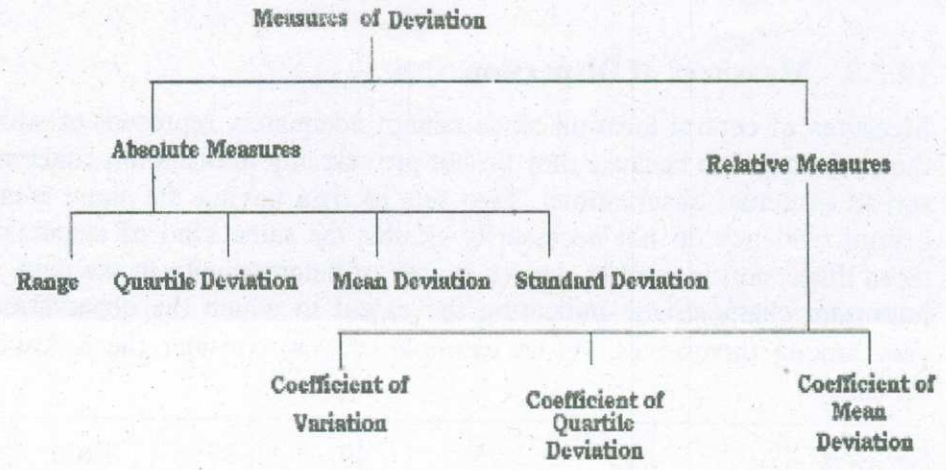
These two series will have the same mean but that does not reflect the character of the data. It is clear from the above example that mean is not sufficient to reveal all the characteristics of data, as both the data set has the same mean but they are significantly different. Suppose these series represent the scores of two batsmen in 5 one-day matches. Though their mean score is the same, the first batsman is much more consistent than the second. Therefore, we require another measure which measures the variability in the data set. These are called the measures of dispersion.

A measure of dispersion is defined as a numerical value explaining the extent to which individual observations vary among themselves. The measures are as following:

1) Absolute Measures of Dispersion
2) Relative Measures of Dispersion

Absolute measures of dispersions measure numerically heterogeneity of data. These measures are not free from unit of measurement. Therefore, these measures cannot be used to measure the degree of heterogeneity between two data sets, which does not have the same unit of measurement. But relative measures are free from unit of measurement and therefore, they are useful in comparing the degree of variability between two sets of data with different unit of measurement.

The following tree diagram gives an overview of measures of dispersions. We will discuss standard deviation in details as this is the most frequently used measure of dispersion.



Absolute Measures of Dispersion

Range: The range of a set of observation is the difference between the maximum and minimum values. In Table 1, the range of number of problems solved = (7 – 3) = 4. Range is very easy to calculate.

Quartile Deviation: Quartile deviation is defined as the half of the difference between the first and third quartiles.

Quartile Deviation: = $(Q_3 - Q_1) / 2$.

Mean Deviation: Mean deviation of a set of observations is the arithmetic mean of the absolute deviations from mean or any other specified value. Here we take absolute deviations so that the positive and negative deviations do not cancel out each other.

Mean deviation about A= $\frac{1}{n}\sum |x_i - A|$, where n is the number of observations

Mean deviation about mean= $\frac{1}{n}\sum |x_i - \bar{x}|$

Standard Deviation: Standard deviation of a set of observations is the square root of the arithmetic mean of squares of deviations from arithmetic mean. Here the deviation of each observation is taken to be the measure of the degree of heterogeneity of data from the central position and these deviations are squared to make all of them a positive number. In such a procedure the positive and negative values do not cancel out each other. After taking the mean of the square of the deviations we take square root of them to get the measure of standard deviation.

Standard deviation is generally denoted by σ and is always is a non negative number. The square of standard deviation (S.D.) is called variance of a variable. S.D. and variance of a variable, say x, is denoted by σ_x and Var(x) or σ^2 respectively.

For ungrouped frequency distribution, the S.D. is given by the following formula:

Standard Deviation = $\sqrt{\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}\right)}$ while for grouped frequency distribution it is given by

$$\text{Standard Deviation} = \sqrt{\left(\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n}\right)},$$

where f_i : frequency of the ith class.

x_i: mid value of the ith class.

Some unique properties of S.D. make it superior to other measures of dispersion. First, it is based on all observations. If the value of one observation changes, the S.D. changes but range and quartile deviation may remain unaffected due to such a change. Secondly, it is least affected by the sampling fluctuation. Thirdly, it is easy to be treated algebraically. Fourthly, given the S.D. of two different groups S.D. along with the number of observations in each group and their mean (arithmetic mean) , the variance for the composite group can be easily calculated using the following formula:

$$\sigma^2 = \{ (n_1 \sigma_1^2 + n_2 \sigma_2^2) + n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2 \} / (n_1 + n_2)$$

where σ^2 : composite variance

n_1 : number of observation in the first group

n_2 : number of observation in the second group

σ_1 : S.D. of the first group

σ_2 : S.D. of the second group

$\bar{x}_1$: mean of the first group

$\bar{x}_2$: mean of the second group

$\bar{x}$: composite mean

Relative Measures of Dispersion: These measures are free of unit of measurement. Generally, the absolute measures are divided by measures of location to arrive at the absolute measure of dispersion. There are three such measures, viz.,

Coefficient of Variation = S.D./ Mean × 100

Coefficient of Quartile Deviation = Quartile Deviation / Median × 100

Coefficient of Mean Deviation = Mean Deviation / Mean or Median × 100

Check Your Progress 3

1) Show that standard deviation is independent change of origin (not explained in the text).

2) If scale measurement is changed, how will it affect the S.D.?

3) Using the fact that S.D. in independent of change of scale, calculate the S.D. of the following data on household size.

Household size	1	2	3	4	5	6	7	8	9
No. of Households	92	49	52	82	102	60	35	24	4

4) There are 50 boys and 40 girls in a class. The average weight of boys and girls are 59.5 and 54 respectively. The S.D. of their weight is 8.38 and 8.23. Find the mean height and composite S.D. of the class.

5) If the S.D. and coefficient of variation of a variable is 4 and 50% respectively, find the mean of the variable.

10.4.3 Measures of Skewness and Kurtosis

Given n observations x_i, i = 1, 2, 3, ...n and an arbitrary constant A for ungrouped frequency distribution

$\frac{1}{n}\sum_{i=1}^{n}(x_i - A)$ is the 1st order moment about A.

$\frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2$ is the second order moment about A

$\frac{1}{n}\sum_{i=1}^{n}(x_i - A)^3$ is the third order moment about A

and so on ..... We will denote them by m_1 , m_2 , m_3 and so on. Again, for grouped frequency distribution

$\frac{1}{n}\sum_{i=1}^{n} f_i(x_i - A)$ is the 1st order moment about A.

$\frac{1}{n}\sum_{i=1}^{n} f_i(x_i - A)^2$ is the second order moment about A

$\frac{1}{n}\sum_{i=1}^{n} f_i(x_i - A)^3$ is the third order moment about A

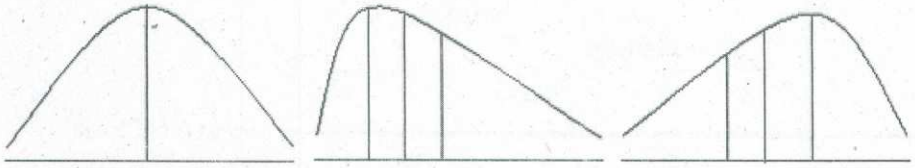
When A=0, we call m_1, m_2 and m_3 as raw moments and when A = $\bar{x}$ we call them central moments and denote them by μ_1, μ_2, μ_3 respectively. You can verify that $\mu_1 = 0$ (first order central moment) and μ_2 (second order central moment)= Var (x).

Skewness:

The frequency distribution of a variable is said to be symmetric if the frequencies of the variable are symmetrically distributed in both sides of the mean. Therefore, if a distribution is symmetric, the values of the variable equidistant from mean will have equal frequencies.

Symmetric distributions are generally bell shaped and mean, median and mode of these distributions coincide. The figures below explain the three types of skewness and their properties in terms of mean, median and mode. The figures show frequency polygon, the values of the variable being measured along the horizontal axis and the frequency for each value of the variable along the vertical axis. There are many methods by which we can measure skewness of a distribution. We discuss these in the following section.

Consider the following figure where Mn implies mean, Md implies median and Mo implies mode.



Symmetric distribution | Positive Skew distribution | Negative Skew Distribution

Pearson's Measures: It is clear from the above figure that if (mean – median) is positive the distribution is positively skewed and if it is negative the distribution is negatively skewed. The more the median and mean are distant the more skew a distribution is. Pearson takes this property of a distribution to derive a measure of skewness.

Pearsonian first measure = (Mean – Mode) / Standard Deviation.

Pearsonian Second Measure = 3 (Mean – Median)/ Standard Deviation.

The measures are relative to S.D. to make them unit free.

Moment Measure: In a symmetric distribution for each positive values of $(x_i - \bar{x})$ there is a corresponding negative value. When these deviations are cubed, they retain their sign. In case of positively skewed distributions

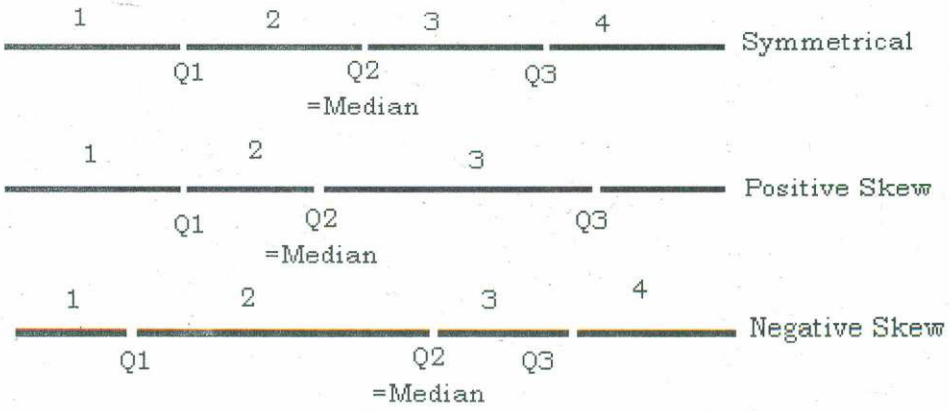
$\sum_{i=i\&(xi-x)>0}^{n} f_i (x_i - \bar{x})^3$ for positive deviations from mean (by deviation from mean we mean $(x_i - x)$)outweighs $\sum_{i=i\&(xi-x)<0}^{n} f_i (xi - \bar{x})^3$ for negative deviations. Note that summing the squares of the deviations from mean makes all the deviations positive and there is no way to infer whether positive deviations are dominated by or dominate negative deviations. Again, summing the deviations from mean makes the summation equal to zero. Therefore, μ_3 is a good measure of skewness. To make it free of unit, we divide it by σ^3.

Moment Measure of Skewness $(\gamma_1) = \mu_3 / \sigma^3 = \mu_3 / (\sqrt{(\mu)^3}$

Bowley's Measures: Bowley's measure of skewness is given by the following formula:

Skewness B = $\{(Q_3 - Q_2) - (Q_2 - Q_1)\}/\{(Q_3 - Q_2) + (Q_2 - Q_1)\}$

$= Q_3 - 2\ Q_2 + Q_1 / Q_3 - Q_1$



No of observations in each section 1, 2 ,3 and 4 are equal

For an exactly symmetrical distribution, Q_2 (Median) lies exactly between Q_1 and Q_2 . For a positively skewed distribution i.e., when the longest tail of the frequency lies to the right, Q_3 will be wider away from Q_2 than Q_1 and vice versa for negatively skewed distribution. The arithmetic mean of the difference between of Q_1 and Q_3 from Q_2 (median) taken relative to quartile deviation gives Bowley's measure of skewness. It is left as an exercise for you to verify.

See that ($Q_1 - Q_2$) is always negative and ($Q_3 - Q_2$) always positive. It is their relative strength which determines the skewness of a distribution.

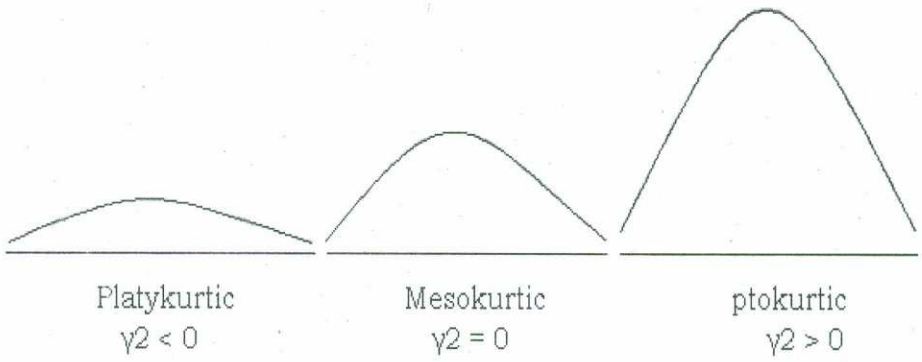
Kurtosis:

Kurtosis refers to the degree of peakedness of the frequency curve. Two distributions having the same average, dispersion and skewness, however, might have different levels concentration of observations near mode. The more dense the observations near the mode, the sharper is the peak of the frequency distribution. This characteristic of frequency distribution is known as kurtosis.

The only measure of kurtosis is based on moments

Kurtosis (γ_2) = (μ_4 / σ^4) – 3 ; where μ_4 is the fourth order central moment

Kurtosis of a distribution could be of three types. The following figures explain them.



Check Your Progress 4

1) Show that first order central moment is always zero.

.....

.....

.....

.....

2) Find the relation between rth order central moment and moment about an arbitrary constant say A.

.....

.....

.....

.....

14.4 LET US SUM UP

The discussion above has shown that the descriptive statistics are used to describe the basic features of the data. They provide simple summaries about the pieces of information collected and their measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Thus, in the process each descriptive statistic reduces lots of data into a simpler summary.

In the first section of this unit, we have discussed the most used techniques of data presentation. Whereas in the later section, we have discussed the tools used for the analytical purpose of the data set. The measures of central tendency, dispersion, skewness and kurtosis are discussed as statistical tools to analyse data.

14.5 KEY WORDS

Arithmetic Mean: A set of realisations of a variable defined as their sum divided by the number of observations. It is usually denoted by $\bar{x}$.

Arithmetic Mean for Ungrouped Data: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; where n is the number of observations.

Arithmetic Mean for Grouped Data: $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i f_i$; where $N = \sum_{i=1}^{n} f_i$.

Bar Diagrams: A group of equispaced rectangular bars, one for each category of given statistical data. All the bars share the common base line. The bar diagrams could be vertical or horizontal.

Class Boundaries: Most extreme values (observations) of a variable which could ever be included in a class interval are called class boundaries.

Class Frequency: Number of observation falling under each class.

Class Limits: Two numbers used to specify the limits of a class interval for the purpose of tallying the original observations.

Class: Classification of observations varying in a wide range into several groups.

Continuous Variable: Variable taking any value within its' range.

Cross-Section Data: Data collected on a single point of time over different sections.

Cumulative Frequency: Number of data elements in a data class and all the previous classes. That is, the sum of all previous frequencies up to the current point.

Coefficient of Mean Deviation: $\frac{M.D}{Median} \times 100.$

Coefficient of Quartile Deviation: $\frac{Q.D}{Median} \times 100.$

Coefficient of Variation: $\frac{S.D}{Mean} \times 100.$

Data: A collection of facts, such as numbers, words, measurements, observations or even just descriptions of things.

Deciles: A method of splitting up a set of ranked data into 10 equally large subsections.

Frequency Density: *Frequency ÷ class width.*

Frequency: How often something happens (usually during a period of time).

Geometric Mean: A special type of average where we multiply the numbers together and then take a square root (for two numbers), cube root (for three numbers) etc. So geometric mean of a set of n observations is nth root of their product.

Harmonic Mean: A kind of average of a set of n numbers by adding the reciprocals of the numbers dividing the sum by n,

The formula: Harmonic Mean = $\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots+\frac{1}{x_n}}$

Grouped Frequency Distribution: Values of the variable in groups or intervals along with the frequencies of the groups or intervals.

Kurtosis: Degree of peakednes of a frequency curve.

Line Diagrams: A graph that uses points connected by lines to show how something changes in value (as time goes by, or as something else happens).

Mean Deviation: Mean of the distances of each value from their mean.

Median: The number in the middle of a given set of numbers arranged in order of increasing magnitude.

Mode: Element that appears most frequently in a given set of elements.

Mid-Point of Class Interval: The value exactly at the middle of a class interval.

Percentiles: Value below which a percentage of data falls.

Pictogram: Use of pictures or symbols to show the value of the data.

Pie Diagrams: A circular chart divided into sectors, each sector showing the relative size of each value.

Qualitative Data or Attributes: Data approximate or characterise but do not measure the attributes, characteristics, properties, etc., of a thing or phenomenon.

Quantitative Data or Variables: Data with numerical information (numbers).

Quartile Deviation: Partial variance between the upper quartile and lower quartile in a distribution given as $Q.D = \frac{Q_3 - Q_1}{2}$.

Quartiles: Values that divide a list of numbers into quarters.

Range: Difference between the lowest and highest values of observation.

Relative Frequency: Ratio of the actual number of favourable events to the total possible number of events. That is share of a frequency of a class in total frequency.

Standard Deviation: A measure of how spread out numbers are.

Skewness: A measure of the degree of asymmetry of a distribution.

Simple Frequency Distribution: A set of scores and their frequency listed in data distribution (how many times each one occurs).

Tabulation: Logical and systematic organisation of statistical data in rows and columns, designed to simplify the presentation and facilitate quick comparison.

Time Series Data: Data collected over a period of time.

Width of a Class: Difference between the upper and lower class boundaries.

14.6 SOME USEFUL BOOKS

Das.N.G. (1996), *Statistical Methods*, M.Das & Co.(Calcutta)

A.M. Goon, M.K.Gupta and B.Dasgupta (2005): *Fundamentals of Statistics*, Vol.I,8th Ed., World Press, Kolkatta

14.7 ANSWER OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1)

- Statistical tables are concise, and because data are systematically grouped and arranged, explanatory matter is minimal.
- Data are more easily read, understood and compared into rows and columns. The reader can understand and interpret a great bulk of data rapidly because he can see significant relationships of data once.
- Tables give the whole information even without combining numerals with textual matter. This is so because tables are so constructed that the ideas they convey can be understood even without reading their textual presentation.

2)

Dist	2004				2005			
	under 18 yrs		18 yrs and above		under 18 yrs		18 yrs and above	
	Male	Female	Male	Female	Male	Female	Male	Female
1								
2								
3								
4								
5								
6								
7								

3) Do it yourself after reading Sub-section 14.3.3.

4) Calculations for drawing pie chart are provided. Draw the pie chart using a protractor. Round figures in the last column up to one decimal places.

Country	Exports of cotton in bales	share in degrees in the Pie Chart
U.S.A	6367	192.5180581
India	2999	90.68032925
Egypt	1688	51.03981186
Brazil	650	19.65395599
Argentina	202	6.107844784
Total	11906	360

Check Your Progress 2

1) Mean = 3.76.

Median is the value of the cumulative frequency corresponding to (N + 1)/2, which is 4.

Mode is value of the variable corresponding to the highest frequency, i.e., 4.

2) Use the formula, Mean – Mode = 3 (Mean – Median)

3) When describing discrete categorical data, mode is preferred because the greatest frequency of responses is important for describing categorical data. Also see the discussion of Section 10.4.1.

4) Do it yourself using the discussion of Section 10.4.1.

Check Your Progress 3

1) Change of origin means shifting the point from which the variable is measured. Let x be a variable after shifting the origin by 'a' units the new variable will be (x – a). You have been asked to show that S.D. of x and (x – a) are same.

S.D. of (x – a) = $\sqrt{\{\sum_{i=1}^{n}((x_i - a) - (\bar{x} - a))^2\}/n}$

where ($\bar{x}$ – a) is the arithmetic mean of the variable (x – a).

S.D. of (x – a) = $\sqrt{\{\sum_{i=1}^{n}((x_i - a) - (\bar{x} - a))^2\}/n}$

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

= S.D. of x. .....(proved)

2) Changing the scale means changing the units of measurement. Let x be a variable after shifting the scale by 'b' the new variable will be x/b.

S.D. of x/b = $\sqrt{\{\sum_{i=1}^{n}(x_i/b - \bar{x}/b)^2\}/n}$

= 1/b × S.D. of x

3) We change the origin by 4 units. The calculations for the derivation of S.D. are shown below. As we have stated earlier Standard Deviation = $\sqrt{\{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2\}/n}$. This formula could be simplified to

$$S.D.^2 = \{\sum_{i=1}^{n} f_i x_i^2\}/\sum_{i=1}^{n} f_i - \{\sum_{i=1}^{n} f_i x_i\}/\sum_{i=1}^{n} f_i$$

Household size x	No. of households f	Y = x - 4	f.y	$f.y^2$
1	92	-3	-276	828
2	49	-2	-98	196
3	52	-1	-52	52
4	82	0	0	0
5	102	1	102	102
6	60	2	120	240
7	35	3	105	315
8	24	4	96	384
9	4	5	20	100
Total	500		17	2217

Using the above formula we get S.D. = 2.11

4) Use the formulae

Composite Mean ($\bar{x}$) = ($n_1 . \bar{x}_1 + n_2 . \bar{x}_2$) / $n_1 + n_2$

$\sigma^2 = \{(n_1 \sigma_1^2 + n_2 \sigma_2^2) + n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2\} / (n_1 + n_2)$

mean = 57.06, S.D. = 8.75

5) Use the formula

Coefficient of Variation = S.D./ Mean × 100

Mean = 8

Check Your Progress 4

1) Do it yourself after reading Section 10.4.3.

2) Expand $\left(x_i - \bar{x}\right)^r = \mu^r$ using binomial theorem $\{(a-b)^r = \sum_{i=0}^{r} {}^rC_i a^i b^{r-i}\}$.

14.8 EXERCISES

1) What is a histogram and how is it constructed? Draw the histogram for the following hypothetical frequency distribution.

class interval	frequency
141-150	5
151-160	16
161-170	56
171-180	19
181-190	4

2) What is pie chart? When is it used?

3) Do you think mean is superior to other measures of central location? If yes, then why? If not, then which one is the best measure of central location and why?

4) From the following table find the missing frequencies a, b given A.M. is 67.45.

height	frequency
60 - 62	5
63 - 65	18
66 - 68	a
69 - 71	b
72 - 74	8
Total	100

5) From the following cumulative frequency distribution of marks obtained by 22 students of IGNOU in a paper, find the arithmetic mean, median and mode.

marks	frequency
Below 10	3
Below 20	8
Below 30	17
Below 40	20
Below 50	22

6) Compute the A.M., S.D. and mean deviation about median for the following data

Scores	frequency
4--5	4
6--7	10
8--9	20
10--11	15
12--13	8
14--15	3

7) Out of 600 observations 350 has the value 3 and rest take the value 0. Find A.M. of 600 observations together.

8) A multinational company has three units in three countries X, Y and Z. The following table summarizes the salary structure of the company in three countries.

	X	Y	Z
Number of employees	20	25	45
Average monthly salary	305	400	320
S.D. of monthly salary	50	40	55

Find the average and S.D. of monthly salaries of all the 90 employees.

9) What is meant by moment of a distribution? What is the difference between raw and central moments?

10) Find the first four central moments and the values of γ_1 and γ_2 from the following frequency distribution. Comment on the skewness and Kurtosis of the distribution.

x	F
21-24	40
25-28	90
29-32	190
33-36	110
37-40	50
41-44	20

UNIT 15 CORRELATION AND REGRESSION ANALYSIS

Structure

15.0 Objectives

15.1 Introduction

15.2 Bivariate Data and Its Presentation

15.3 Simple Correlation Analysis

- 15.3.1 Meaning, Nature, Assumptions and Limitations
- 15.3.2 Measures of Correlation
  - 15.3.2.1 Scatter Diagram
  - 15.3.2.2 Karl Pearson's Correlation Coefficient
  - 15.3.2.3 Coefficient of Rank Correlation

15.4 Simple Regression Analysis

- 15.4.1 Meaning and Nature
- 15.4.2 Ordinary Least Square Method of Estimation
- 15.4.3 Properties of Linear Regression

15.5 Standard Error of Estimate

15.6 Unexplained Variation and Explained Variation

15.7 Partial and Multiple Correlation and Regression

15.8 Methods of Estimating Non-Linear Equations

15.9 Let Us Sum Up

15.10 Key Words

15.11 Some Useful Books

15.12 Answer or Hints to Check Your Progress

15.13 Exercises

15.0 OBJECTIVES

After going through this unit, you will understand the techniques of correlation and regression. In particular, you will appreciate the concepts like:

- scatter diagram;
- covariance between two variables;
- correlation coefficient;
- least square estimation method of regression; and
- partial and multiple correlation and regression.

15.1 INTRODUCTION

We start with the presentation of bivariate data and proceed to deal with the nature of association between two variables. In the process, we will be exposed to the use of correlation and regression analyses and their applications to a host economic problems.

In everyday life, in business or policymaking we try to find out by how much one variable is getting affected by another variable. Suppose an economy is expected to grow faster. Then we have to find out the factors, which influence growth, and then to control those to lay down policies accordingly. But how do we know which factor affects growth by what degree? Suppose investment, political stability, technical know-how and growth rate of population affect economics growth rate. But how do we know which one of these affects how much.

Similarly, a firm wants to find out how much of its sales are affected by advertisement. Does advertisement of its product increase its sales or not?

In all the above problems we use correlation and regression analyses, which enable us to get a picture of by what degree a variable affects another.

15.2 BIVARIATE DATA AND ITS PRESENTATION

The statistical methods that have been discussed in the previous unit were concerned with the description and analysis of a single variable. In this unit, we intend to discuss the methods, which are employed to determine the relationship between two variables and to express such a relationship numerically or definitely. Data containing information on the two variables simultaneously are called bivariate data. For example, we may have data on the heights and weights of students of a particular University, or data on the amount of rainfall and yield of rice. In rest of the unit we will assume to have a pair of values (x, y) being denoted by (x_i, y_i) $i = 1, 2, 3 \ldots n$.

Table 15.1: Bivariate Data Showing Height and Weight of 10 Students

	Height (inches)	Weight (kgs.)
1	64	60
2	68	65
3	71	78
4	59	57
5	62	60
6	63	66
7	72	76
8	66	69
9	57	58
10	73	80

Table 15.1 shows height and weight of 10 students of Calcutta University. When in bivariate data a large number of pair of observation is available, it becomes necessary to organise the data in form of a two-way frequency table called bivariate frequency correlation table. From such a table, we can derive a univariate frequency table. A univariate frequency distribution, derived from the bivariate frequency distribution for a specified value (or class interval) of the other variable is called conditional distribution.

Table 15.2: Bivariate Frequency Table (showing ages of 70 husbands and wives)

	Age of Wife (in years)							
Age of Husband (in years)		18-23	23-28	28-33	33-38	38-43	43-48	Total
	21-26	3						3
	26-31		6					6
	31-36			9	3			12
	36-41			2	15	1		18
	41-46				4	20		24
	46-51						7	7
	Total	3	6	11	22	21	7	70

The first and last columns and the first and last rows show the univariate frequency distributions of the age of husbands and wives respectively. The following two tables show conditional distribution of ages of husbands when the age of wife is 33 and above but below 38 and conditional distribution of ages of wives when the age of husband is 36 and above but below 41.

Table 3a: Conditional Distribution of Ages of husbands when age of wife is 33-38

Age of husbands	Frequency
21-26	0
26-31	0
31-36	3
36-41	15
41-46	4
46-51	0
Total	22

Table 3b: Conditional Distribution of Ages of wives when age of husband is 36-41

Age of Wife	Frequency
18-23	0
23-28	0
28-33	2
33-38	15
38-43	1
43-48	0
Total	18

The bivariate frequency distribution makes presentation of data easy. Another way of presenting bivariate data is a scatter diagram. In scatter diagram each bivariate observation can be geometrically represented by a point on the graph paper, where the value of one variable is shown along the horizontal axis and that of another along the vertical axis.

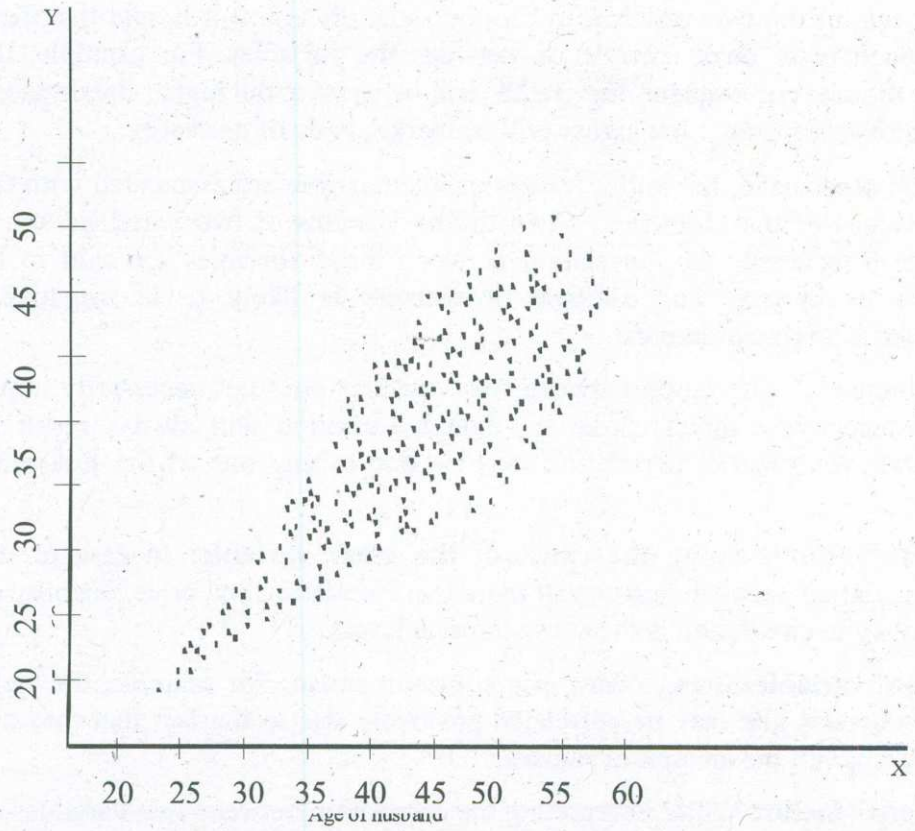


Fig. 15. 1: Scatter Diagram Presenting Bivariate Data of Ages of Husbands and Wives

Check Your Progress 1

1) From the following bivariate data calculate the conditional mean values of y when x = 2, x = 7 and x = 8.

y / x	0	1	2	3	4	5	6
1	1	7	1	2	1	5	9
2	2	11	6	12	3	7	8
3	4	4	13	9	5	10	3
4	6	0	3	2	10	4	0
5	7	8	4	10	11	4	6
6	8	3	2	11	4	12	1
7	7	4	5	1	13	4	4
8	6	0	2	3	2	1	8

15.3 SIMPLE CORRELATION ANALYSIS

15.3.1 Meaning, Nature, Assumption and Limitations

Bivariate data set may reveal some kind of association between two variables x and y and we may be interested in numerically measuring the degree of strength of this association. Such a measure can be performed with correlation. For example, we want to measure the degree of association between rainfall and yield of rice. Are they positively related, i.e., high value of rainfall is associated with high value of yield of rice or are they negatively related or does there not exist any relationship between them? If higher values of the one variable are associated with higher values of the other or when lower values of the one are accompanied by the lower values of the other (in other words,

movements of the two variables are in the same direction) it is said that there exists positive or direct correlation between the variables. For example, the greater the sides of a square, the greater will be its area; the higher the dividend declared by a company, the higher will be market price of its shares.

If on the other hand, the higher values of one variable are associated with the lower values of the other (i.e., when the movements of two variables are in opposite directions), the correlation between those variables are said to be negative or inverse. For example, investment is likely to be negatively correlated with rate of interest.

The presence of correlation between two variables does not necessarily imply the existence of a direct causation, though causation will always result in correlation. In general, correlation may be due to any one of the following factors:

i) **One variable being the cause of the other variable:** In case of the association between quantity of money in circulation and price, quantity of money in circulation is the cause of price levels.

ii) **Both variables being result of a common cause:** For example, the yield of rice and jute may be correlated positively due to the fact that they are related with the amount of rainfall.

iii) **Chance factor:** While interpreting the correlation between two variables, it is essential to see if there is any likelihood of the relationship. It might sometimes happen that between two variables a fair degree of correlation may be observed but there is no likelihood of any relationship between them. For example, wholesale price index of India and average height of its male population.

Between two variables, the degree of association may range all the way from no relationship at all to a relationship so close that one variable is a function of the other. Thus, correlation may be:

1) Perfectly positive
2) Limited positive degree
3) No correlation at all
4) Limited negative degree
5) Perfectly negative

When we find a perfect positive relation between two variables, we designate it as +1. In case of perfect negative we describe it as –1. Thus, correlation between any two variables must vary between –1 and +1.

Correlation may be linear or non-linear. If the amount of change in one variable tends have a constant ratio to the amount of change in the other, then the correlation is said to be linear. Here we will study linear correlation only. This is often called simple correlation.

Limitations of Simple Correlation

Simple correlation analysis deals with two variables only and it explores the extent of linear relationship between them (if x and y are linearly related, then we can write $y = a + bx$). But as we have noted earlier correlation between two variables may be due to the fact that they are affected by a third variable. Simple correlation analysis may not give the true nature of association between two variables in such an event. Ideally, one should take out the effect of the 3rd

variable on the first two and then go on measuring the strength of association between them. But this is not possible under simple correlation analysis. In such situations, we use partial and multiple correlations, which will be discussed later.

In simple correlation analysis, we assume linear relationship between two variables but there may exist non-linear relationship between them. In that case, simple correlation measure fails to capture the association.

Again, strong relationship (linear) between two variables will imply that correlation between them is high (either stark positive or stark negative) but the converse is not necessarily true.

15.3.2 Measures of Correlation

We use following methods to measure simple correlation between two variables:

1) Scatter Diagram
2) Karl Pearson's Coefficient of Correlation
3) Coefficient of Rank Correlation

15.3.2.1 The Scatter Diagrams

Scatter diagrams help to visualise the relationship between two variables.

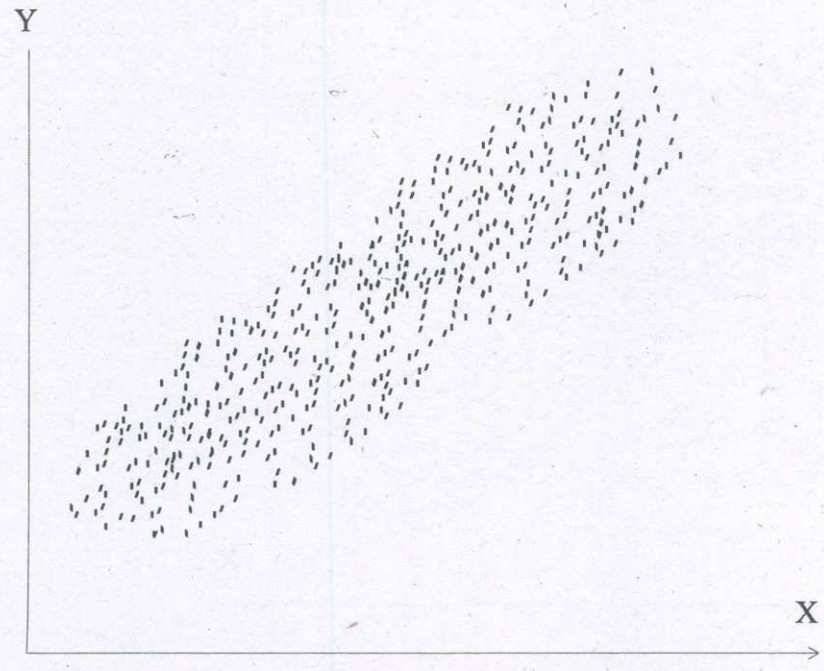


Fig. 15.2(A): Positive Correlation

The way in which points on the scatter diagram lie, indicate the nature of relationship between two variables. From scatter diagram, we do not get any numerical measurement of correlation. If the path formed by the dots starts from lower left hand comes to upper right hand corner (Figure 15.2A), it means there exists a positive correlation. If the dots form a straight line from lower left hand corner to upper right hand corner, then there is exact positive correlation (+1) between the variables (Figure 15.2B).

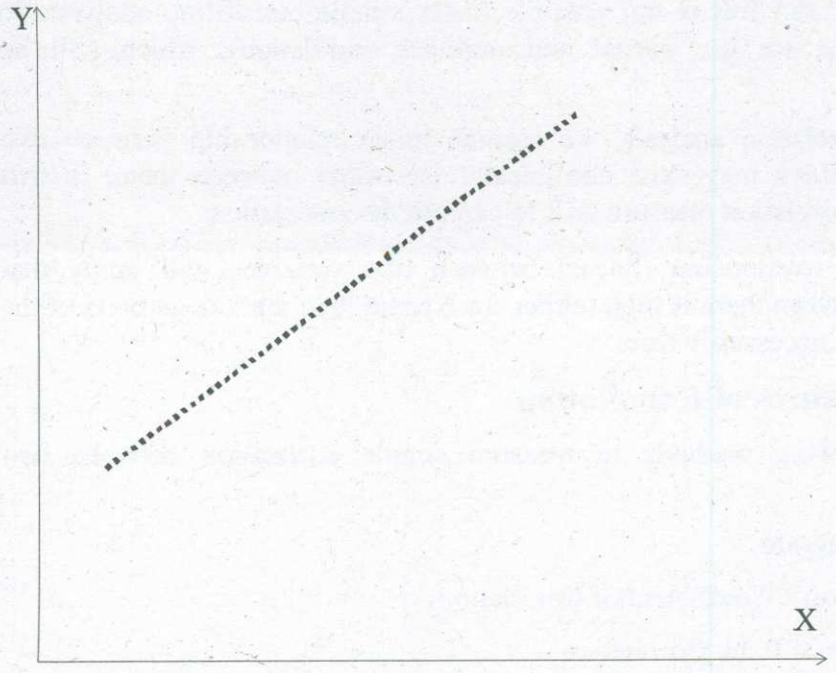


Fig. 15.2(B): Exact Positive Correlation

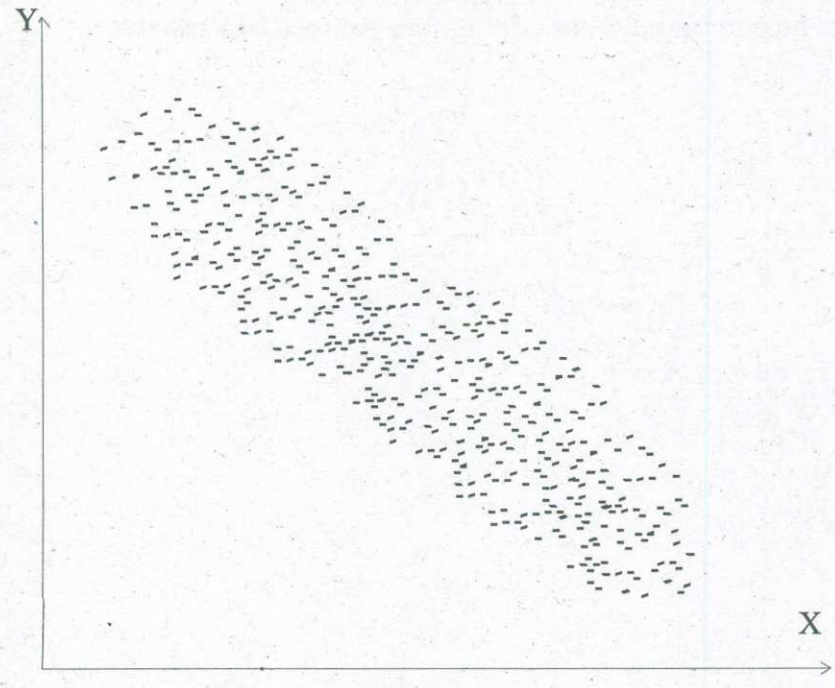


Fig. 15.2 (C): Negative Correlation

If on the other hand, the path starts from the upper left hand corner and ends at lower right hand corner, then there exists negative correlation (Figure 15.2C) and if the dots lie on a straight line in the same fashion, then there exists exact negative (–1) correlation between the variables (Figure 15.2D). But if the path formed by the dots does not have any clear direction, then there is no correlation or spurious correlation at all between the two variables (Figure 15.2 E and F).

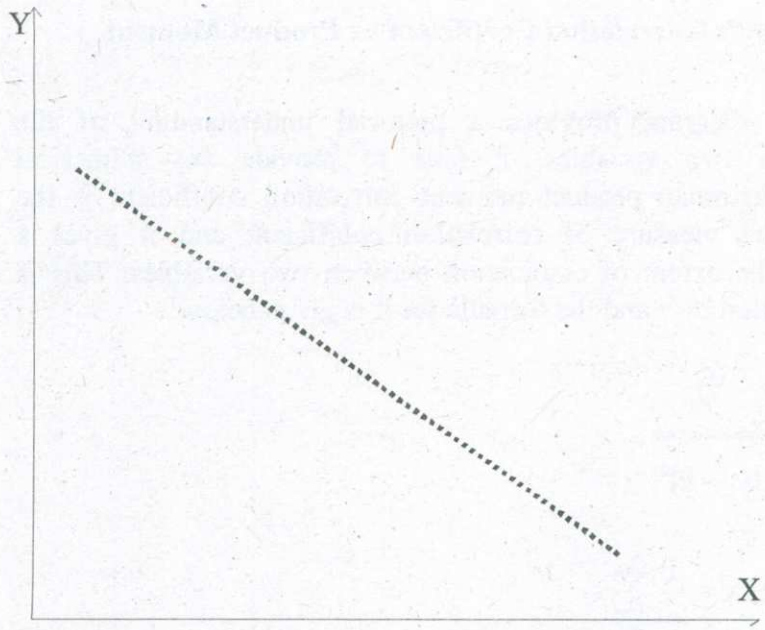


Fig. 15.2(D): Exact Negative Correlation

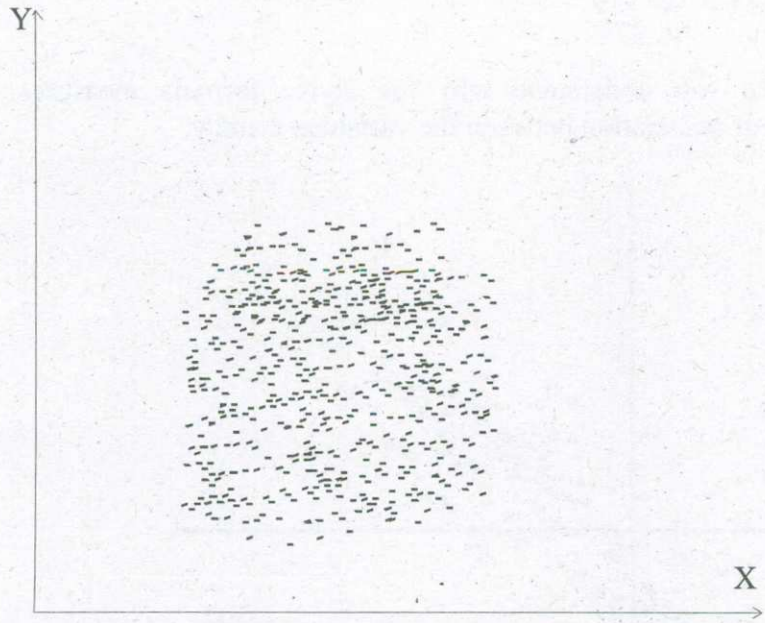


Fig. 15.2(E): Zero Correlation

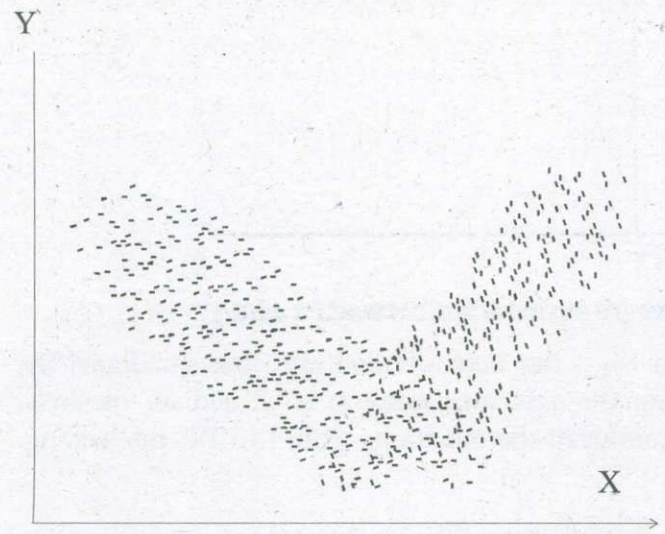


Fig. 15.2(F): Zero Correlation

15.3.2.2 Karl Pearson's Correlation Coefficient or Product Moment Correlation

Although a scatter diagram provides a pictorial understanding of the relationship between two variables, it fails to provide any numerical relationship. The Pearsonian product moment correlation coefficient is the most commonly used measure of correlation coefficient and it gives a numerical value of the extent of association between two variables. This is symbolically represented by γ and the formula for it is given below:

$$\Upsilon = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 . \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ = mean of $x = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\bar{y}$ = mean of $y = \frac{1}{n}\sum_{i=1}^{n} y_i$

Figure 14.3 will help you understand why the above formula measures effectively the degree of association between the variables x and y.

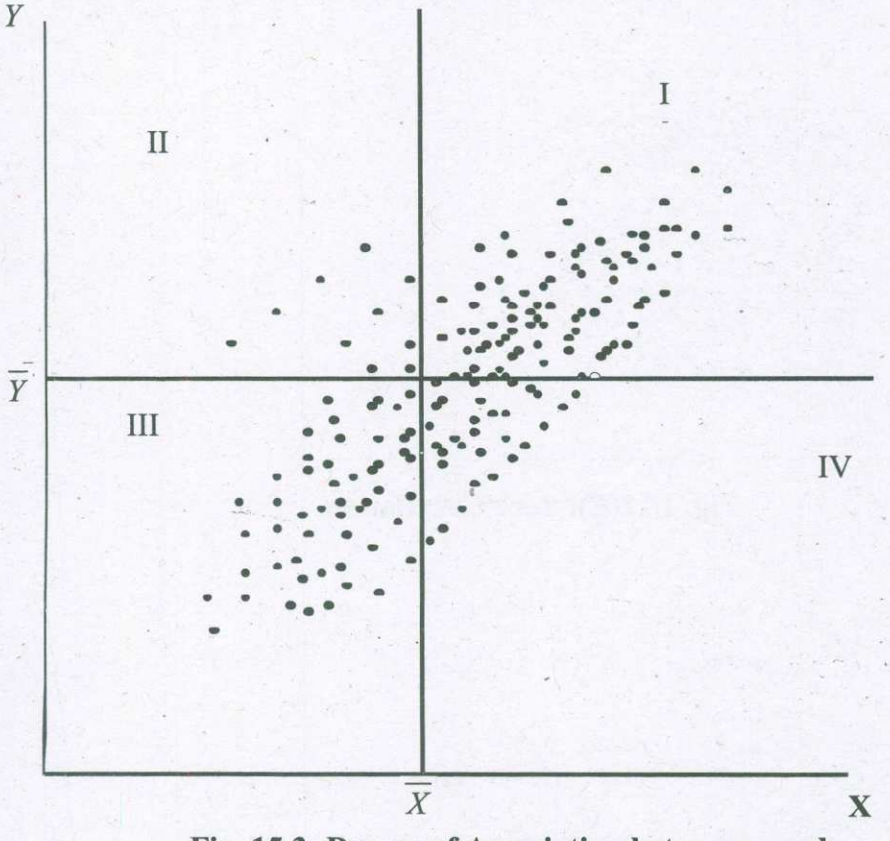


Fig. 15.3: Degree of Association between x and y

The scatter diagram in Figure 14. 3 has been divided into four quadrants by drawing two perpendiculars on the axis measuring x at $\bar{x}$ and on the axis measuring y at $\bar{y}$. We have numbered the quadrants from I to IV, proceeding anticlockwise.

Notice in the numerator of the formula for γ that we have $(x_i - \bar{x})$ and $(y_i - \bar{y})$. These measure the deviations of values of the variable x and y from their

means. Points lying in quadrant I have high values of x as well as high values of y. Therefore, for these points $(x_i - \bar{x})$ and $(y_i - \bar{y})$ scores are both positive. Again, for points lying on the quadrant III, both x and y take low values. Therefore, both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ scores for this region is negative. Thus, for all points laying in quadrant I and III, $(x_i - \bar{x})(y_i - \bar{y})$ is positive. Notice the more points lie in these two region the association between them is positive in nature.

Similarly, for points lying in quadrant II, $(x_i - \bar{x})$ is negative, whereas $(y_i - \bar{y})$ scores are positive. While for points lying in quadrant IV $(x_i - \bar{x})$ scores are positive, $(y_i - \bar{y})$ scores lying there are negative. Therefore, for all points lying in quadrant II and IV, $(x_i - \bar{x})(y_i - \bar{y})$ term is negative. Note that the more points lie in these two regions, the association between x and y is negative. Consequently, for all points lying in quadrant II and IV, $(x_i - \bar{x})(y_i - \bar{y})$ term is negative.

Thus, if $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is positive, then relatively more points are there in quadrants I and III than in quadrants II and IV and there is a positive association between the variables and vice versa. Mean of $(x_i - \bar{x})(y_i - \bar{y})$ scores is called covariance between x and y. This is denoted by cov(x, y). So,

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \text{ which can be simplified as } = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\bar{y}.$$

The cov(x,y) is a measure of association between x and y, which is independent of sample size [as we divide $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ by n to get cov(x,y)] but it is not free from units of x and y. To make it unit free, we divide it by standard deviation of x (σ_x) and standard deviation of y (σ_y). As we know,

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Thus, we get Pearson's Product moment correlation coefficient, which is free from units as well as from sample size and write:

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\gamma = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Properties of γ

i) The correlation coefficient γ is independent of the choice of both origin and scale. This means, if u and v are two new variables defined as:

$$u = \frac{x-c}{d} \qquad v = \frac{y-c'}{d'}$$

where c, d, c′ and d′ are arbitrary constants, then correlation coefficient between u and v (γ_{uv}) will be same as correlation coefficient between x and y (γ_{xy}), i.e., $\gamma_{uv} = \gamma_{xy}$.

ii) The correlation coefficient (γ) is a pure number and it is free from units.

iii) The correlation coefficient lies between +1 and –1.

Proof:

Let x and y be two variables and we have n pairs of observation (x_1y_1), (x_2,y_2), …, (x_n,y_n) on them. Their mean and standard deviations are respectively $\bar{x}$, $\bar{y}$ and σ_x, σ_y.

We define two new variables u and v, where

$$u_i = \frac{x_i - \bar{x}}{\sigma_x} \qquad v_i = \frac{y_i - \bar{y}}{\sigma_x}$$

for all i = 1, 2, …, n

(This process of demeaning a variable and then dividing it by its standard deviation is called standardisation.)

$$\sum_{i=1}^{n} u_i^2 = \frac{1}{\sigma_x^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n\sigma_x^2}{\sigma_x^2} = n$$

$$\sum_{i=1}^{n} v_i^2 = \frac{1}{\sigma_y^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{n\sigma_y^2}{\sigma_y^2} = n$$

Now,

$$\sum_{i=1}^{n}(u_i + v_i)^2 \geq 0$$

$$\text{or, } \sum_{i=1}^{n} u_i^2 + \sum_{i=1}^{n} v_i^2 + 2\sum_{i=1}^{n} u_i v_i \geq 0$$

$$\text{or, } n + n + 2.\gamma n \geq 0$$

$$\text{or, } \gamma \geq -1 \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Again,

$$\sum_{i=1}^{n}(u_i - v_i)^2 \geq 0$$

$$\text{or, } \sum_{i=1}^{n} u_i^2 + \sum_{i=1}^{n} v_i^2 - 2\sum_{i=1}^{n} u_i v_i \geq 0$$

$$\text{or, } n + n - 2.\gamma n \geq 0$$

$$\text{or, } 1 \geq \gamma \qquad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Thus, from (1) and (2) we get, $-1 \leq \gamma \leq 1$.

15.3.2.3 Coefficient of Rank Correlation

The Karl Pearson's product moment correlation coefficient cannot be used in cases where the direct quantitative measurement of the variables is not possible. (For example, consider honesty, efficiency, intelligence, etc.). However, we can rank the different items and apply the Spearman's method of rank differences for finding the degree of correlation.

Suppose we want to measure the extent of correlation between ranks obtained by a group of 10 students in Economics and Statistics. Since we do not have actual marks (or the ranks are not quantitative variables), we will use Pearson's Rank Correlation Coefficient. It is often denoted by ρ (read as 'rho') and is given by the formula.

$$\rho = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$$

where D is the absolute difference between the ranks of an individual, n is the number of pairs ranked.

Here the variables x and y take values 1 to n (i.e., natural members). As we know from our previous unit, mean and standard deviation of first n natural numbers are given by $\frac{n+1}{2}$ and $\sqrt{\frac{n^2-1}{12}}$ respectively.

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}$$

$$= \frac{1}{n}\frac{\left\{\sum x_i^2 + \sum y_i^2 - \sum (x_i - y_i)^2\right\}}{2} - \bar{x}\bar{y}$$

$$= \frac{1}{n}\frac{2.n(n+1)(2n+1)}{6 \times 2} - \frac{\sum_{i=1}^{n} D_i^2}{2n} - \left(\frac{n+1}{2}\right)^2$$

[Let $x_i - y_i = D_i$]

$\because \sum x_i^2 = \sum y_i^2$ = sum of squares of natural numbers $= \frac{n(n+1)(2n+1)}{6}$

$$= \left\{\frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}\right\} - \frac{\sum_{i=1}^{n} D_i^2}{2n}$$

$$= \frac{(n^2-1)}{12} - \frac{\sum_{i=1}^{n} D_i^2}{2n}$$

or, $\gamma_{xy} = \frac{\text{cov}(x,y)}{\sigma_x . \sigma_y}$

Here $\sigma_x = \sigma_y = \sqrt{\frac{n^2-1}{12}}$

$$\therefore \rho = \frac{(x^2-1)}{12} - \frac{\sum_{i=1}^{n} D_i^2}{2n} \Bigg/ \sqrt{\frac{n^2-1}{12}} \times \sqrt{\frac{n^2-1}{12}}$$

$$\therefore \rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

In the calculation of ρ (rank correlation coefficient) if several individuals have the same score, it is called the case of 'tied ranks'. The usual way to deal with such cases is to allot average ranks to each of these individuals and then calculate product moment correlation coefficient. The other way is to modify the formula for ρ as

$$\therefore \rho' = \frac{1 - 6\left\{\sum_{i=1}^{n} D_i^2 + \sum (t^3 - t)/12\right\}}{n(n^2-1)}$$

where t is the number of individuals involved in a tie, no matter whether in first or second variable.

Properties of ρ

The rank correlation coefficient lies between –1 and +1. When the ranks of each individual in the two attributes (e.g., rank in Statistics and Economics) are equal, ρ will take the value 1. When the ranks in one attribute is just the opposite of the other (say, the student who topped in Statistics got lowest marks in Economics and so on), ρ will take the value –1.

Check Your Progress 2

1) Calculate the product moment correlation coefficient between heights and weights of 10 students in Table 1 and comment on its value.

.....

.....

.....

.....

.....

.....

2) Calculate the product moment correlation coefficient between Age of Husbands and Age of Wives from the data in Table 2.

.....

.....

.....

.....

.....

.....

3) Show that the formula for cov(x, y) can be simplified as

$$\text{cov}(x,y) = \frac{\sum x_i y_i}{n} - \bar{x}.\bar{y}$$

.....

.....

.....

.....

.....

4) Show that γ is independent of change of origin and scale.

.....

.....

.....

.....

5) In a drawing competition 10 candidates were judged by 2 judges and the ranks given by them are as follows:

Candidate	A	B	C	D	E	F	G	H	I	J	K
Ranks by Judge	1	4	8	6	7	1	3	2	5	10	9
Ranks by Judge	2	3	9	6	5	1	2	4	7	8	10

Compute the coefficient of rank correlation.

15.4 SIMPLE REGRESSION ANALYSIS

15.4.1 Meaning and Nature

So far we have discussed the method of computing the degree of correlation existing between two given variables. In bivariate data, we may have one variable of particular interest (dependent variable) and the other variable is studied for its ability of explaining the former. In such situations, we would like to guess the definite relationship between the two variables. This is the idea of regression. A line will have to be fitted to the points plotted in the scatter diagram to calculate the amount of change that will take place in the dependent variable (generally, denoted by y) for a unit change in the explanatory variable or the independent variable (denoted by x). Equation of such a line is called regression line (here for the time being we will restrict ourselves to linear regression only). We can make predictions of the dependent variable (y) for a particular value of the independent variable (x), by estimating the regression line of y on x. In the rest of the section we will try to deduce the most efficient way to derive the regression lines.

The term 'regression line' was first used by Sir Francis Galton in describing his findings of the study of hereditary characteristics. He found that the height of descendants has a tendency to depend on (regress to) the average height of the race. Such a tendency led Galton to call the 'line of average relationship' as the

'line of regression'. Nowadays the term 'line of regression' is commonly used even in business and economic statistics to describe the line of average relationship.

15.4.2 Ordinary Least Square Method of Estimation

The standard form of the linear regression of Y on X is given by

$Y = a + bX$

where a and b are constants. The first constant 'a' is the value of Y when X takes the value 0. The constant 'b' indicates the slope of the regression line and gives us a measure of the change in Y values due to a unit change in X. This is also called the regression coefficient of Y on X and is denoted as b_{yx}. If we know a and b, then we can predict the values of Y for a given values of X. But in the process of making that prediction we might commit some error. For example, in the diagram when $X = X_i$, Y takes the value Y_i, but our regression line Y on X predicts the value Y_i. Here e_i is the magnitude of error we make in predicting the dependent variable. We shall choose the values of 'a' and 'b' in such a fashion that these errors (e_i's) are minimised. Suppose there are n pairs of observations (y_i, x_i), $i = 1, 2, \ldots, n$. Then if we want to fit a line of the form

$\hat{Y}_i = a + bX_i$ (Regression Line Y on X)

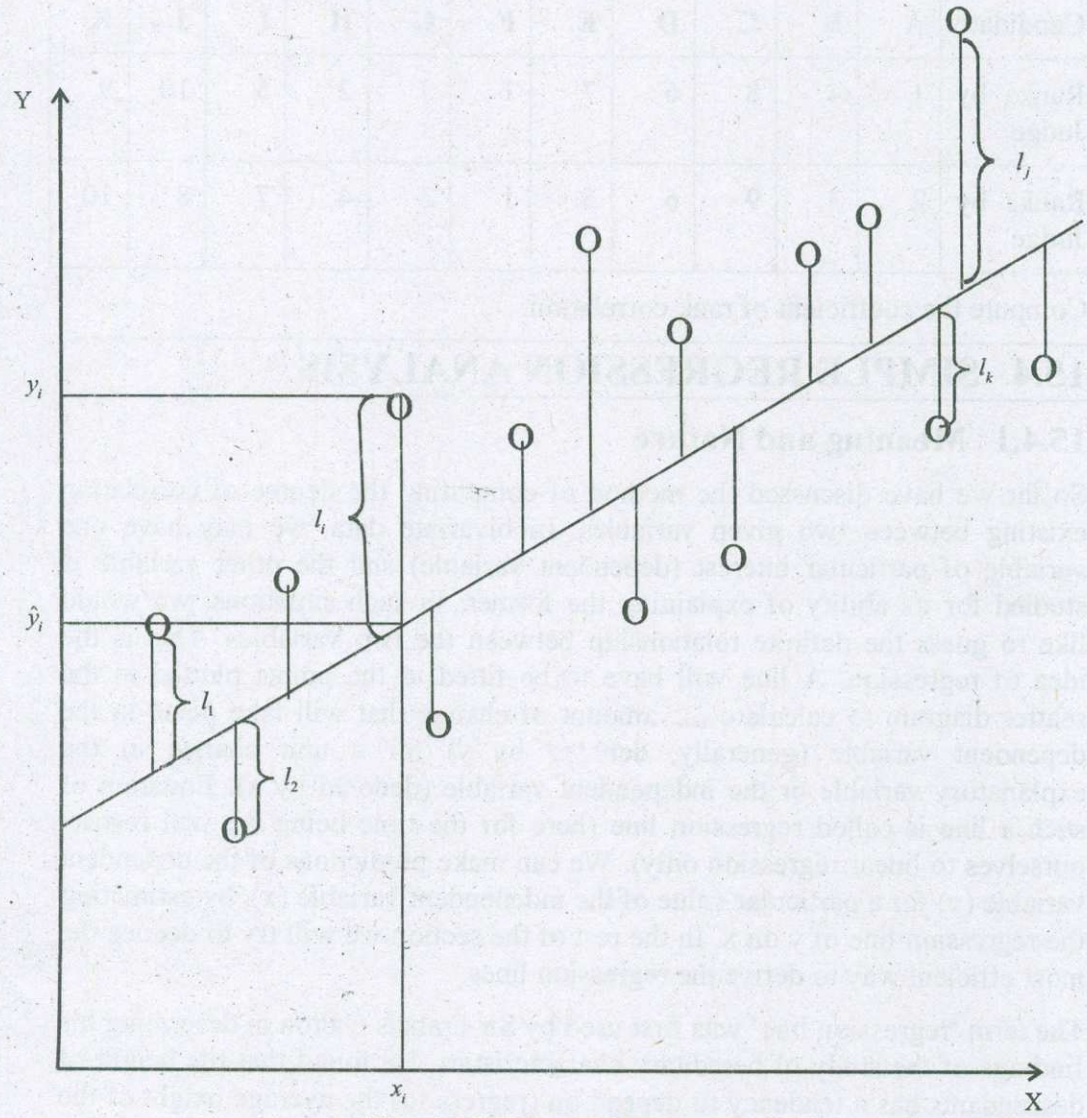


Fig. 15.4: Regression Lines

then for every X_i, i = 1, 2, …, n, the regression line (Y on X) will predict $\hat{Y}_i$ (the predicted value of the variable Y). Therefore, the measure of error of prediction is given by

$e_i = Y_i - \hat{Y}_i$ (See figure 15.4)

Note that e_i could be positive as well as negative. To get the total amount of error we make while filling a regression line we cannot simply sum the e_i's. For the reason that positive and negative e_i's will cancel out each other and will reduce the total amount of error. Therefore, we take the sum of the squares of e_i's and we take the sum of the squares of e_i's to minimise $\left(\sum_{i=1}^{n} e_i^2\right)$ and choose 'a' and 'b' to minimise this amount. This process of obtaining the regression lines is called Ordinary Least Square (OLS) method. In deriving equation Y on X we assume that the values of X are known exactly and those of Y are subject to error.

$e_i = Y_i - \hat{Y}_i$

$e_i = Y_i - \hat{a} - \hat{b}X_i$ [$\hat{a}$ and $\hat{b}$ are the estimated values of a and b]

$$\therefore \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2$$

We minimise $\sum_{i=1}^{n} e_i^2$ with respect to $\hat{a}$ and $\hat{b}$ and first order conditions,

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{a}} = -\sum_{i=1}^{n} 2(Y_i - \hat{a} - \hat{b}X_i) = 0$$

or, $$\sum_{i=1}^{n} 2(Y_i - \hat{a} - \hat{b}X_i) = 0$$

or, $$n\hat{a} + \hat{b}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\frac{\partial \sum_{i=1}^{n} e_i}{\partial \hat{b}} = -\sum_{i=1}^{n} 2X(Y_i - \hat{a} - \hat{b}X_i) = 0$$

or, $$\hat{a}\sum_{i=1}^{n} X_i + \hat{b}\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

(Check whether second order conditions are satisfied or not)

Equation (1) and (2) are called normal equations.

From equation (1), $$\hat{a} = \frac{1}{x}\sum_{i=1}^{n} Y_i - b\frac{1}{x}\sum_{i=1}^{n} X_i$$

$$= \bar{Y} - b\bar{X}$$

Substituting '$\hat{a}$' in Equation (2),

$$(\bar{Y} - b\bar{X})\sum_{i=1}^{n} X_i + \hat{b}\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

or, $$\hat{b}\left(\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} X_i Y_i - \bar{Y}\sum_{i=1}^{n} X_i$$

or, $$\hat{b}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = \sum_{i=1}^{n} X_i Y_i - n\bar{Y}.\bar{X} \quad \left[\because \sum_{i=1}^{n} X_i = n\bar{X}\right]$$

$$\text{or,} \quad \hat{b} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2}$$

$$= \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\text{cov}(X,Y)}{\sigma_x.\sigma_y}.\frac{\sigma_y}{\sigma_x} = \gamma.\frac{\sigma_y}{\sigma_x}$$

$$\therefore \hat{a} = \bar{Y} - \frac{\text{cov}(X,Y)}{\text{var}(X)}.\bar{X}$$

Thus, the regression equation of Y on X is given by

$$\hat{Y}_i = \bar{Y} - \frac{\text{cov}(X,Y)}{\text{var}(X)}.\bar{X} + \frac{\text{cov}(X,Y)}{\text{var}(X)}.X_i$$

$$\text{or} \quad \hat{Y}_i - \bar{Y} = \frac{\text{cov}(X,Y)}{\text{var}(X)}(X_i - \bar{X})$$

Similarly, the regression equation of X on Y is of the form

$$X_i = a + bY_i$$

which we get using OLS method and in deriving the equation X on Y. We assume values of Y are known and that of X are subject to errors.

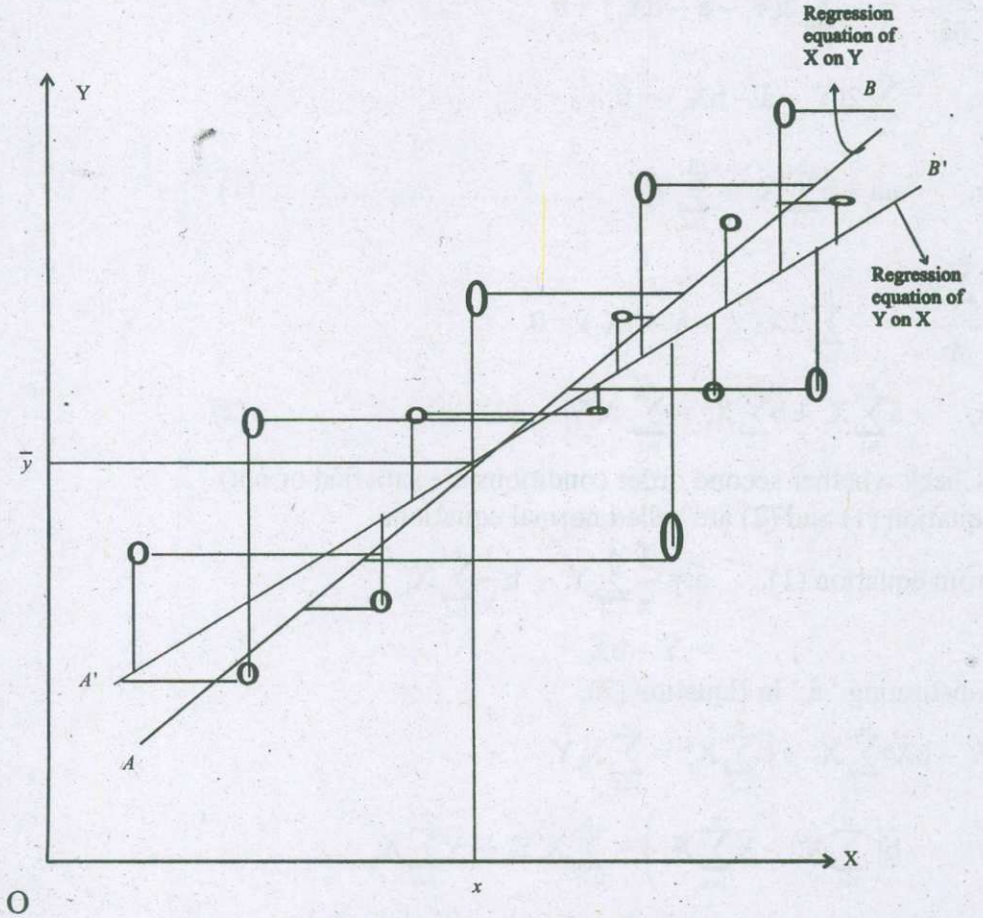


Fig. 15.5: Regression Lines

In Figure 15.5, A′B′ is the regression line of Y on X whereas AB is the regression line of X on Y. A′B′ is obtained by minimising the vertical errors and ab is obtained by minimising the horizontal errors.

The lines intersect at $(\bar{X},\bar{Y})$. The intercept as well as the slope of the two lines are different.

The equation of regression equation X on Y is given below:

$$\hat{X}_i - \bar{X} = \frac{cov(X,Y)}{var(Y)}(Y_i - \bar{Y})$$

15.4.3 Properties of Linear Regression

Let us define two regression lines as

$\hat{Y}_i - \bar{Y} = b_{yx}(X_i - \bar{X})$ [Y on X] and

$\hat{X}_i - \bar{X} = b_{xy}(Y_i - \bar{Y})$ [X on Y]

b_{xy} and b_{yx} are called coefficients of regression.

Note the following properties:

1) If b_{yx} and b_{xy} denote the slopes of the regression lines Y on X and X on Y, respectively, i.e., $b_{yx} \times b_{xy} = \gamma^2$

   i.e., product of coefficients of regression is equal to square of the correlation coefficient.

2) $b_{yx} = \gamma.\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = \gamma.\frac{\sigma_x}{\sigma_y}$

3) γ, b_{xy} and b_{yx} all have the same sign. If γ is zero then b_{xy} and b_{yx} are zero.

4) The angle between the regression lines depends on the correlation coefficient (γ). If $\gamma = 0$, they are perpendicular. If $\gamma = +1$ or -1 they coincide. As γ increases numerically from 0 to 1 or -1, angle between the regression lines starts diminishing from 90^o to 0^o.

Check Your Progress 3

1) You are given that the variance of X is 9. The regression equations are $8X - 10Y + 66 = 0$ and $40Y - 18Y = 214$. Find

   i) average values of X and Y

   ii) γ_{xy}

   iii) σ_y

2) Regression of savings (s) of a family on income may be expressed as $s = a + \frac{y}{m}$ where a and m are constants. In a random sample of 100 families the variance of savings is one-quarter of variance of incomes and the correlation is found to be 0.4 between them. Obtain the value of m.

3) The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 men.

	X	Y
Mean	53	142
Variance	130	165

Find the appropriate regression equation and use it to estimate the blood pressure of a man whose age is 45.

15.5 STANDARD ERROR OF ESTIMATE

In the above analysis we showed that the linear regression analysis enables us to predict or estimate the value of the dependent variable for any value of the independent variable. But our estimate of the dependent variable, not necessarily, would be equal to the observed data. In other words, the regression line may not pass through all the points in the scatter diagram.

Suppose, we fit a regression line of yield of rice on the amount of rainfall. But this regression line will not enable us to make estimates exactly equal to the observed value of the yield of rice when there is a certain amount of rainfall. Thus, we may conclude that there is some error in the estimate. The error is due to the fact that yield of crop is determined by many factors and rainfall is just one of them. The deviation of the estimated or predicted value from the observed value is due to influence of other factors on yield of rice.

In order to know, how far the regression equation has been able to explain the variations in Y, it is necessary to measure the scatter of the points around the regression line. If all the points on the scatter diagram fall on the regression line, it means that the regression line gives us perfect estimates of the values of Y. In other words, the variations in Y are fully explained by the variations in X and there is no error in the estimates. This will be the case when there is perfect correlation between X and Y ($Y = +1$ or -1). But if the plotted points do not fall upon the regression line and scatter widely from it, the use of regression equation as an explanation of the variation in Y may be questioned. The regression equation will be considered useful if in estimating values of Y only if the estimates obtained by using it are more correct than those made without it. Then only, we can be sure of the functional relationship between X and Y.

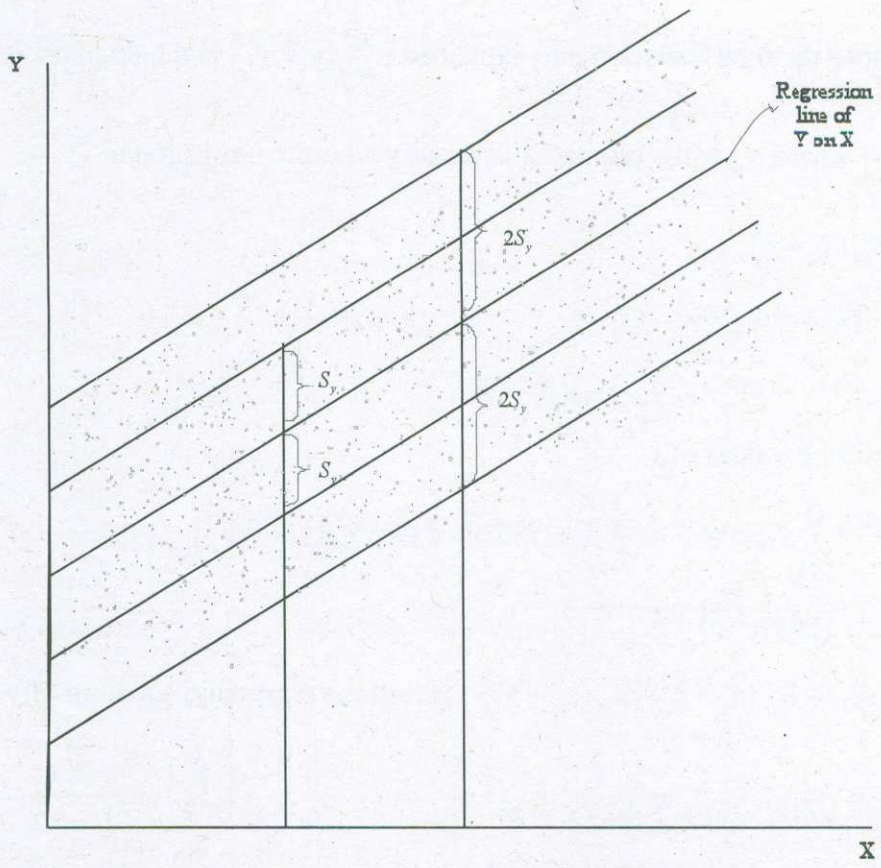


Fig. 15.6: Regression Line and Errors

If the measure of the scatter of the points from the regression line is less than the measure of the scatter of the observed values of Y from their mean, then we can infer that the regression equation is useful in estimating Y. The scatter of points from the regression equation is called 'the standard error in estimating Y'. It is obtained commonly by the following formula:

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

The interpretation of the standard error of estimate (S_y) is the same as that of the standard deviation of univariate frequency distribution. As in the case of normal frequency distribution 68.27% and 95.45% of the observation lie in the interval of (mean $\pm 1.\sigma$) and (mean $\pm 2.\sigma$) respectively, in case of standard error the same percent of observations lie in the area formed by the two parallel lines in each side of the regression line at a distance of $1.S_y$ and $2.S_y$ measured along Y axis respectively. (see Figure 15.6).

15.6 UNEXPLAINED VARIATION AND EXPLAINED VARIATION

For a set of pair of observations (x_i, y_i), $i = 1, 2, \ldots, n$, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is called the total variation in y, where $\bar{y}$ is the arithmetic mean of the variable y. We

can decompose the total variation into explained $\left(\sum_{i=1}^{n}(\hat{y}_i\bar{y}_i)^2\right)$ and unexplained $\left(\sum_{i=1}^{n}(\hat{y}_i\bar{y}_i)^2\right)$ where $\hat{y}_i$ is the estimated value of y where $x - x_i$ is given.

$\therefore\ \hat{y} = \bar{y} + b_{yx}(x_i - \bar{x})$

$(y_i - \bar{y})^2 = \{(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\}^2$

$= (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$

Summing over all values of i

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Now, $\sum(y_i - \bar{y}_i)(\hat{y}_i - \bar{y})$

$= \sum\{(y_i - \bar{y}_i) - b_{yx}(x_i - \bar{x})\}\{b_{yx}(x_i - \bar{x})\}$ [From the regression equation of y on x

$= \sum b_{yx}(y_i - \bar{y}_i)(x_i - \bar{x}) - \sum b_{yx}^2(x_i - \bar{x})^2$

$= b_{yx}\sum(y_i - \bar{y}_i)(x_i - \bar{x}) - b_{yx}^2\sum(x_i - \bar{x})^2$

$= b_{yx}.n.cov(x,y) - b_{yx}^2.n\sigma_x^2$

$$\left[\text{Since } \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = cov(x,y) \text{ and } \frac{1}{x}\sum(x_i - \bar{x})^2 = \sigma_x^2\right]$$

$= nb_{yx}.cov(x,y) - n.b_{yx}(\sigma_x^2.b_{yx})$

$= n.b_{yx}.cov(x,y) - n.b_{yx}.\gamma.\frac{\sigma_Y}{\sigma_X}.\sigma_x^2$

$= n.b_{yx}.cov(x,y) - n.b_{yx}.\frac{cov(x,y)}{\sigma_x.\sigma_y}.\sigma_y.\sigma_x = 0$. Thus,

$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$+ \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
Total variation =(Unexplained variation) + (Explained variation)		
Or		
Total sum of square (TSS)= Explained Sum of Square (ESS) + Residual Sum of Square (RSS)		

The expression equation estimates explains only $\hat{y}_i$ portion of the actual value of y_i. The rest of y_i, i.e., $(\hat{y}_i - \bar{y})$ is unexplained or often termed as residual. Hence, $\sum(y - \hat{y}_i)$ is called unexplained variation.

It can be shown that,

$$\frac{ESS}{TSS} = \gamma^2 \quad \text{or} \quad 1 - \frac{RSS}{TSS} = \gamma^2$$

i.e., proportion of total variation explained by regression = γ^2. Thus, when

$\gamma = \pm 1$; ESS = TSS or RSS = 0

$\gamma = 0$; RSS = TSS or ESS = 0

15.7 PARTIAL AND MULTIPLE CORRELATION AND REGRESSION

So far in simple correlation and regression analysis we studies the strength of association between two variables as well as the specific form (linear) of relationship between them. The above analysis was based on the unrealistic assumption that one variable is influenced by another variable only. But this is not true always. The yield of rice depends not only on rainfall but also on the amounts of fertiliser and pesticides, temperature and many more other factors. Again the weight of a student crucially depends on his/her height as well as diet and chest measurement, etc. In all the above cases, we are concerned with three or more variables simultaneously. These types of distributions are called multivariate distribution.

The measure of the extent of combined influence of a group of variables on another variable is the concern of multiple correlation whereas extent of association between two variables after eliminating the effect of other variables is called partial correlation.

Multiple regression

In multiple regression we try to predict the value of one variable given the values of other variables. Let us consider the case of three variables y, x_1 and x_2. We assume there exists linear relationship between them. Thus,

$y = a + bx_1 + cx_2$

where, a, b and c are constants.

We apply the same method of OLS to obtain the estimates ($\hat{a}$, $\hat{b}$ and $\hat{c}$) of a, b and c to minimise the sum of the square of errors.

Thus, our task is to

$$\underset{\hat{a}, \hat{b}, \hat{c}}{\text{Min}}\, E = \sum e_i^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_{1i} - \hat{c}.x_{2i})^2$$

Differentiating E with respect to $\hat{a}$, $\hat{b}$ and $\hat{c}$ we get following three *normal* equations:

$$\sum (y_i - \hat{a} - \hat{b}x_{1i} - \hat{c}.x_{2i}) = 0 \quad \ldots\ldots\ldots\

or, $\hat{a} = (\bar{y} - \hat{b}\bar{x}_1 - \hat{c}.\bar{x}_2)$

Substituting $\hat{a} = (\bar{y} - \hat{b}\bar{x}_1 - \hat{c}.\bar{x}_2)$ in equations (2) and (3), we get

$$\sum y_i x_{1i} = (\bar{y} - \hat{b}x_1 - \hat{c}.x_2)\sum x_{1i} + \hat{b}\sum x_{1i}^2 + \hat{c}\sum x_{2i}x_{1i} \quad \ldots\ldots\ldots\ldots(4)$$

$$\sum y_i x_{2i} = (\bar{y} - \hat{b}x_1 - \hat{c}.x_2)\sum x_{2i} + \hat{b}\sum x_{2i}.x_{1i} + \hat{c}\sum x_{2i}^2 \quad \ldots\ldots\ldots\ldots(5)$$

From (4)

$$\sum y_i x_{1i} - x\bar{y}.\bar{x} = \hat{b}\left(\sum x_{1i}^2 - x\bar{x}_1^2\ \hat{b}\sum x_{1i}^2\right) + \hat{c}\left(\sum x_{2i}.x_{1i} - x\bar{x}_1\bar{x}_2\right)$$

Dividing both sides by n, we get

$$\text{cov}(y,x_1) = \sigma_1^2.\hat{b} + \text{cov}(x_1,x_2). \quad \ldots\ldots\ldots\ldots(6)$$

Similarly, from Equation (5), we get

$$\text{cov}(y,x_2) = \text{cov}(x_1,x_2).\hat{b} + \sigma_{x_2}^2.\hat{c} \quad \ldots\ldots\ldots\ldots(7)$$

Solving (5) and (6), we get

$$\hat{b} = \frac{\text{cov}(y,x_1)\sigma_{x_2}^2 - \text{cov}(x_1,x_2).\text{cov}(y,x_2)}{\sigma_{x_1}^2\sigma_{x_2}^2 - \text{cov}(x_1,x_2)^2}$$

$$= \frac{\sigma_{x_1}^2\sigma_{x_2}^2\left\{\frac{\sigma_y}{\sigma_{x_1}}\frac{\text{cov}(y,x_1)}{\sigma_{x_1}\sigma_y} - \frac{\sigma_y}{\sigma_{x_2}}\frac{\text{cov}(x_1,x_2)}{\sigma_{x_1}\sigma_{x_2}}\cdot\frac{\text{cov}(y,x_2)}{\sigma_y.\sigma_{x_2}}\right\}}{\sigma_{x_1}^2\sigma_{x_2}^2\left\{1 - \left(\frac{\text{cov}(x_1x_2}{\sigma_{x_1}.\sigma_{x_2}}\right)^2\right\}}$$

$$= \frac{\frac{\sigma_y}{\sigma_{x_1}}\left(\gamma_{yx_1} - \gamma_{x_1x_2}.\gamma_{yx_2}\right)}{1 - \gamma_{x_1x_2}^2}$$

[γ_{xy} = correlation coefficient between variables x and y]

and

$$\hat{c} = \frac{\sigma_{x_1}^2.\text{cov}(y,x_2) - \text{cov}(x_1,x_2).\text{cov}(y,x_2)}{\sigma_{x_1}^2\sigma_{x_2}^2 - \text{cov}(x_1,x_2)^2}$$

which can be further simplified as

$$\hat{c} = \frac{\frac{\sigma_y}{\sigma_{x_2}}\left(\gamma_{yx_2} - \gamma_{yx_1}.\gamma_{x_1x_2}\right)}{1 - \left(\gamma_{x_1x_2}\right)^2}$$

Note: $\hat{b}$ and $\hat{c}$ give the effect of x_1 and x_2 on y respectively.

Since $\hat{b}$ is the per unit effect of x_1 on y after eliminating the effects of x_2, it gives the partial regression coefficient of y and x_1 eliminating the effects of x_2. It is often denoted by $b_{12.3}$. Similarly, $\hat{e}$ is often denoted by $b_{13.2}$.

The general multiple linear regression takes of the form:

$$Y_i = B_1X_{1i} + B_2X_{2i} + B_3X_{3i} + ... + + B_kX_{ki} + u_i \qquad i = 1, 2, ..., n$$

where u_i is the error term.

A detailed discussion of the above form of regression equation will be taken up in the Econometric Course of the MA (Economics) Programme and not pursued here. For the present, we solve the coefficient vector (B_1, B_2, ..., B_k) applying the same philosophy of ordinary least squares.

Partial and Multiple Correlation

When we have data on more than two variables simultaneously, correlation between two variables may be of two types, viz.,

i) Partial correlation

ii) Multiple correlation

While measuring partial correlation, we eliminate the effect of other variables on the two variables we are measuring correlation between.

In case of multiple correlation, we measure the product moment correlation coefficient between the observed values of a variable and the estimated values of that variable from a multiple linear regression.

i) **Partial Correlation**

Suppose we have data on three variables y, x_1 and x_2. We assume they posses a linear relationship among them specified by,

$y_i = a + bx_{1i} + cx_{2i}$

To obtain the partial correlation between y and x_1 we have to eliminate the effect of x_2 from both of them. Then the product moment correlation coefficient between the residuals (values of y and x_1 after the effect of x_2 has been eliminated from them) gives the partial correlation.

Let us consider the bivariate regression between y and x_2, and x_1 and x_2 as,

$y = \alpha + \beta_{02}x_2$

$x_1 = \alpha' + \beta_{12}x_2$

But as we have shown earlier that x_2 might not be able to explain variations in y and x_1 fully. We eliminate the effects of x_2 from both of them as follows:

$$e_{yi} = y_i - (\hat{\alpha} + \hat{B}_{02}.x_{2i}) \qquad i = 1, 2, ..., n$$

$$e_{x_{1i}} = x_i - (\hat{\alpha} + \hat{B}_{12}.x_{2i}) \qquad i = 1, 2, ..., n$$

The product moment correlation coefficient between e_{yi} and e_{x1} is partial correlation between y and x_1. This is given by the following formula,

$$\gamma_{yx_1.x_2} = \frac{\gamma_{yx_1} - \gamma_{yx_2}.\gamma_{x_2x_3}}{\sqrt{\left(1-\gamma^2_{yx_2}\right)\left(1-\gamma^2_{x_1x_2}\right)}}, \text{ where}$$

$\gamma_{yx_1.x_2}$ is read as partial correlation between y and x_1 eliminating the effect of x_2.

Partial correlation coefficient always lies between –1 and +1.

ii) **Multiple Correlation Coefficient**

The product moment correlation coefficient between y_i and $\hat{y}_i (= \hat{a} + \hat{b}x_{1i} + \hat{c}x_{2i})$ gives the multiple correlation coefficient.

The multiple correlation coefficient y on x_1 and x_2 is given by

$$R_{y.x_1x_2} = \frac{\sqrt{\gamma^2_{yx_1} + \gamma^2_{y.x_2} - 2\gamma_{yx_1}.\gamma_{x_1x_2}.\gamma_{y.x_2}}}{\sqrt{1-\gamma^2_{x_1x_2}}}$$

Multiple correlation coefficient is always considered to be positive.

Check Your Progress 4

1) Given the following coefficients:

$\gamma_2 = 0.41$

$\gamma_{13} = 0.71$

$\gamma_{23} = 0.50$

find $\gamma_{12.3}$, $\gamma_{13.2}$ and $R_{1.23}$ where the symbols have their usual significance.

.....

.....

.....

.....

.....

15.8 METHODS OF ESTIMATION OF NON-LINEAR EQUATIONS

All the measures of relationships and dependence of one variable on others discussed earlier are captured only when the relationship is linear. But in practice, mostly, relationships are non-linear. Such relationships may be parabolic, exponential and geometric. We adopt different techniques to estimate them.

i) **Parabolic Relationship**

Suppose the relationship between two variable is

$Y = a + bx + cx^2$

We have data on x pairs of observations $(x_i\ y_i)$; $i = 1, 2, \ldots, n$.

Using the method of least squares, the constants a, b, c could be estimated by solving the following 3 equations, which we get in the same method as used earlier.

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

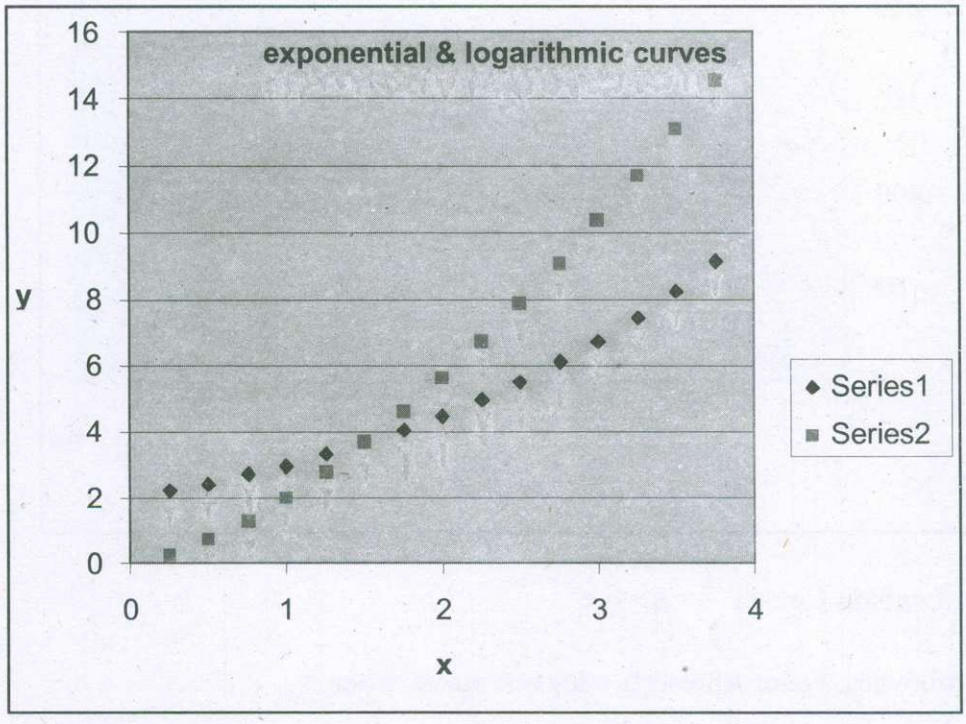
In practical problem if values of x are too high, we often introduce the method of change of origin of the independent variable (i.e., we deduct a constant number from the values of x, which does not affect the results) or from both x and y to ease the calculations.

ii) **Exponential and Geometric Curves**

Take the equations of following form:

$y_i = a.b^{x_i}$ (exponential form)

$y_i = a.\ x_i^{b}$ (logarithmic form)



For a=2, b=1.5 and x=(.25, .5, .75,1,..... 3.75), Series 1 represents exponential curve and Series 2 represents logarithmic curve.

If we take logarithms of both sides of the exponential and logarithmic forms of equation given above, we get

$\log y_i = \log a + \log bx_i$ and

$\log y_i = \log a + b.\log x_i$.

If we assume

$y'_i = \log y_i$, $\log a = A$

$x'_i = \log x_i$, $\log b = B$,

then the equations reduce to simple bivariate equation of the form

$y'_i = A + B.x_i$

$y'_i = A + b.x'_i$.

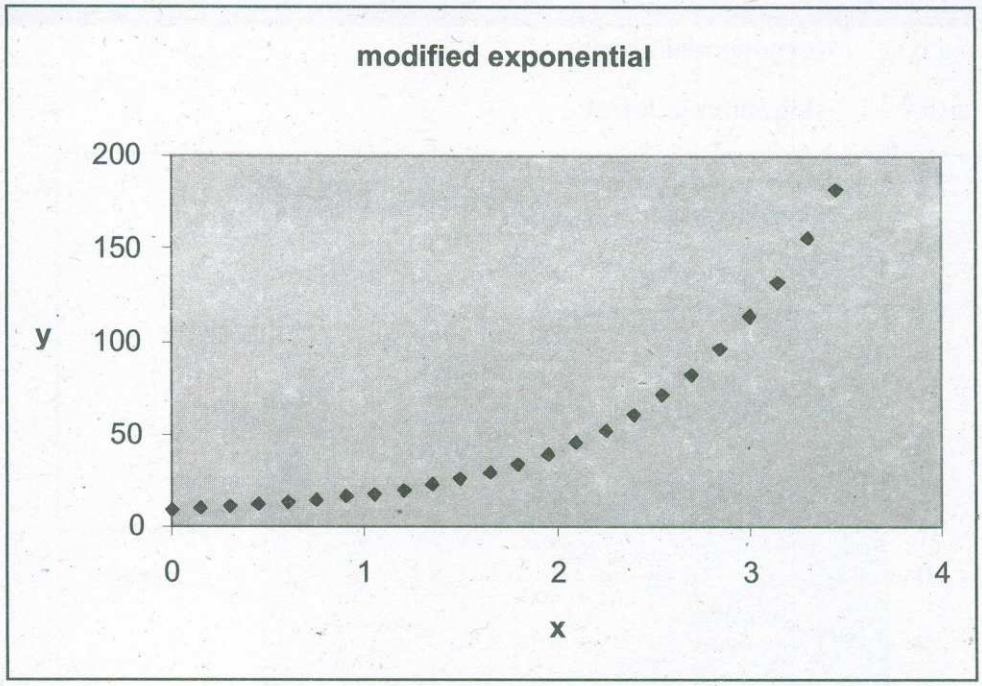
Estimating the coefficients using OLS is easy from these equations. After getting the estimates of the coefficients of transformed equation, we can get back the original coefficients by simple algebraic manipulations.

Other than parabolic, exponential and geometric relationship two variables may show relationship, which is best, filled by the following curves:

1) **Modified Exponential Curve: $y = a + b.c^x$**

The following figure represents a modified exponential curve, when

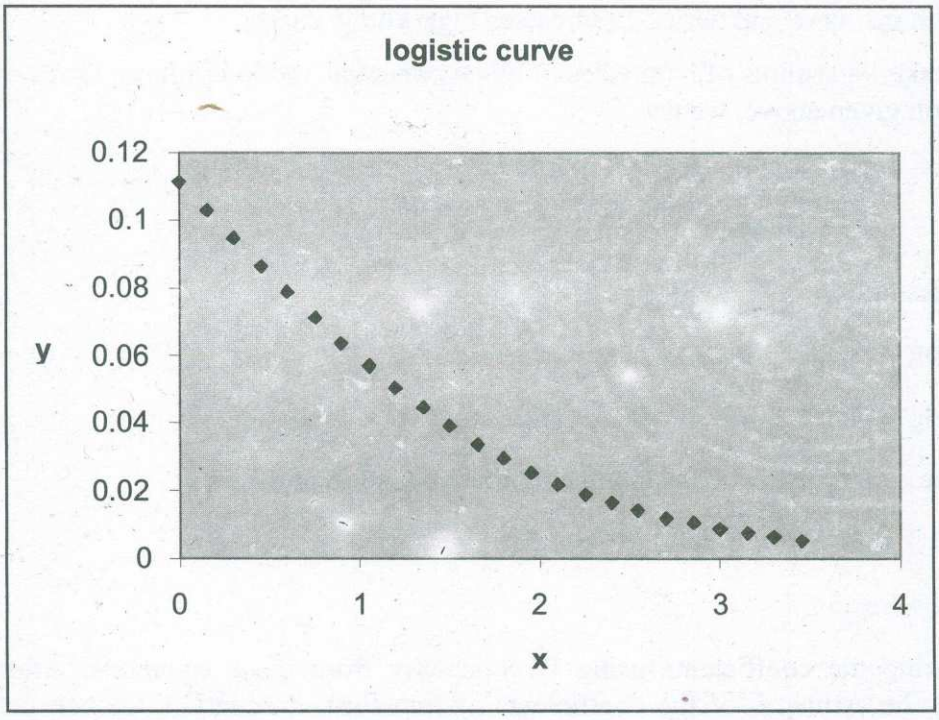
A=5, b=4, c=3, x = (1 2 3 .....25)



2) **Logistic Curve: $\frac{1}{y} = a + b.c^x$**

The following figure represents a logistic curve, when

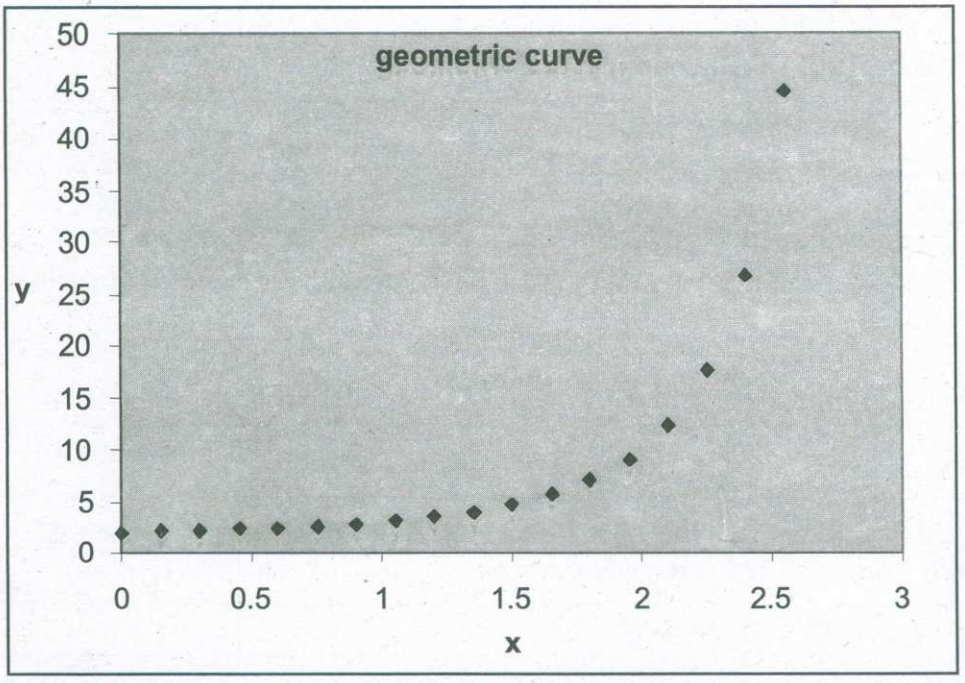
A=5, b=4, c=3, x =(0, 1, 2, 3, ...25)



3) Geometric Curve: $\log y = a + bc^x$

The following figure represents a geometric curve when

$a = 0.5, b = 0.2, c = 3, x = (0, .15, .30, .60, \ldots\ldots, 2.7)$



Note that all the above curves can be fitted using suitable methods and we have not included in the present discussion.

Examples

1) Fit a second degree parabola to the following data:

x	y
1	2.18
2	2.44
3	2.78
4	3.25
5	3.83

Solution

Let the curve be $y = a + bx + cx^2$

Fill up the table yourself:

	x	y	x^2	x^3	x^4	xy	x^2y
	1						
	2						
	3						
	4						
	5						
TOTAL	15						

Construct the normal equations as in (14.8) using the data from the table.

Solve the three equations and check

$Y = 2.048 + .081x + .055x^2$

Check Your Progress 5

1) Fit an equation of the form $y = ab^x$ to the following data

X	2	3	4	5	6
Y	144	172.8	207.4	248.8	298.5

2) Fit a parabola of 2nd degree ($y = a + bx + cx^2$) to the following data using OLS.

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

Find out the difference between the actual value of y and the value of y obtained from the fitted curve when $x = 2$.

15.9 LET US SUM UP

The main objective of many economic studies is to either reveal the cause and effect of a certain economic phenomenon or to forecast the values that would be taken by some economic variables given the values of some variables which are most likely to affect it. In this unit, we first described bivariate data and discussed the techniques to analyse the relationship between those two variables. Simple correlation analysis is useful to determine the nature or the strength of relationship between two variables, while the simple regression analysis helps recognise the exact functional relationship between two variables. Things do get complicated when there are more than two variables. Here we have introduced the concepts of partial and multiple correlation and multiple regressions, which are the tools of multivariate data analysis. The unit closes discussion after an introduction of the forms of some important curves such as parabolic, exponential and geometric often used and can be estimated through OLS method.

15.10 KEY WORDS

Bivariate data: Data relating to two variables is called bivariate data.

Correlation Coefficient: The degree of association between two variables is measured by correlation coefficient. Positive correlation between two variables implies high (low) values of one variable are accompanied high (low) values of the other. Similarly, negative correlation between two variables implies high (low) values of one variable are accompanied low (high) values of the other. The formula of correlation between two variables is given by

$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Scatter Diagram: The diagram we obtain after simply plotting bivariate data, where the axes measure two variables.

Coefficient of Rank Correlation: If the two variables in a bivariate data are ranks of different individuals, say, ranks of the students of a class in the subjects Mathematics and History, we obtain a measurement of association between these two ranks through Spearman's rank correlation coefficient, which is given by the following formula $\rho = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$.

Simple Regression Equation of X on Y: The simple regression equation of X on Y is defined as follows:

$\hat{X}_i = \bar{X} - \frac{\text{cov}(X,Y)}{\text{var}(Y)} \times \bar{Y} + \frac{\text{cov}(X,Y)}{\text{var}(Y)} \times Y_i$. The regression equation of X on Y gives the estimated values of X given the value of the variable Y.

Simple Regression Equation of Y on X: The simple regression equation of Y on X is defined as follows:

$\hat{Y}_i = \bar{Y} - \frac{\text{cov(X,Y)}}{\text{var (X)}} . \bar{X} + \frac{\text{cov(X,Y)}}{\text{var (X)}} . X_i$. The regression equation of Y on X gives the estimated values of Y given the value of the variable X.

Standard Error of an Estimate: The standard deviation of the parameters are called the standard error of an estimate and given by $S_y = \sqrt{\frac{1}{n}\sum\left(Y_i - \hat{Y}\right)^2}$.

Partial Correlation Coefficient: Suppose we have multivariate (more than two variable) data. The correlation coefficient between two variables after eliminating the effect of the other variables from both of the variable gives the partial correlation coefficient.

Multiple Correlation Coefficient: The product moment correlation coefficient between the observed values of a variable and the estimated values of that variable is called multiple correlation coefficient.

15.11 SOME USEFUL BOOKS

Das.N.G. (1996), *Statistical Methods*, M.Das & Co.(Calcutta).

Freund J.E. (2001), *Mathematical Statistics*, Prentice Hall of India.

Goon A.M., Gupta M.K., Dasgupta B. (1991), *Fundamentals of Statistics*, Vol.1, World Press, Calcutta.

Hoel, P (1962), *Introduction to Mathematical Statistics*, Wiley John & Sons, New York.

15.12 ANSWER OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) Conditional distribution of y when x = 2

$\sum fy = 49$

$\sum y.fy = 154$

$(\bar{y}/x = 2) = \frac{154}{49} = 3.14$

y	fy	y.fy
0	2	0
1	11	11
2	6	12
3	12	36
4	3	12
5	7	35
6	8	48

Conditional distribution of y when x = 7.

$\sum fy = 38$

$\sum y.fy = 113$

$(\bar{y}/x = 7) = \frac{113}{38} = 2.97$

y	fy	y.fy
0	7	0
1	4	4
2	5	10
3	1	3
4	13	52
5	4	20
6	4	24

Conditional distribution of y when x = 8.

$\sum fy = 22$

$\sum y.fy = 74$

$(\bar{y}/x = 8) = \frac{74}{22} = 3.36$

Y	fy	y.fy
0	6	0
1	0	0
2	2	4
3	3	9
4	2	8
5	1	5
6	8	48

Check Your Progress 2

1)

i	Height (h_i)	Weight (w_i)	$h_i - \bar{h}$	$w_i - \bar{w}$	$(h_i - \bar{h})^2$	$(w_i - \bar{w})^2$	$(h_i - \bar{h})(w_i - \bar{w})$
1	64	60	–1.5	–0.9	2.25	0.81	1.35
2	68	65	2.5	4.1	6.25	16.81	10.25
3	71	78	5.5	17.1	30.25	292.41	94.05
4	59	57	–6.5	–3.9	42.25	15.21	25.35
5	62	60	–3.5	–0.9	12.25	0.81	3.15
6	63	6	–2.5	–54.9	6.25	3014.01	137.25
7	72	76	6.5	15.1	42.25	228.61	98.15
8	66	69	0.5	8.1	0.25	65.61	4.05
9	57	58	–8.5	–2.9	72.25	8.41	24.65
10	73	80	7.5	19.1	56.25	364.81	143.25
Total	655	609					
Mean	65.5	60.9					

$$\gamma = \frac{\sum_{i=1}^{n}(h_i - \bar{h})(w_i - \bar{w})}{\sqrt{\{\sum(h_i - \bar{h})^2\}\{\sum(w_i - \bar{w})^2\}}}$$

$$\frac{541.5}{\sqrt{270.5}\ \sqrt{4006.9}} = \frac{541.5}{16.45 \times 63.30} = \frac{541.5}{1041.08} = .52$$

2)

		Age of Wife (in years)						
Age of Husband (in years)		18-23	23-28	28-33	33-38	38-43	43-48	Total
	21-26	3						3
	26-31		6					6
	31-36			9	3			12
	36-41			2	15	1		18
	41-46				4	20		24
	46-51						7	7
	Total	3	6	11	22	21	7	70

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{2720}{70} = 38.85$$

x_i	f_i	$x_i - \bar{x}$
23.5	3	–15.35
28.5	6	–10.35
33.5	12	–5.35

$\bar{y} = 35.71$

Similarly, calculate $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i f_i}{\sum_{i=1}^{n} f_i}$

$$\text{cov}(x,y) = \sum_{i=1}^{b} \frac{\sum_{i=1}^{b} (x_i - \bar{x})(y_i - \bar{y}).f_y}{N}$$

$$= \frac{(-15.35)(-15.21)^3 + (-10.35)(-10.21)6}{70}$$

f_y : frequency when $x = x_i$ and $y = y_i$

Calculate σ_x and σ_y

Then obtain $\gamma = \dfrac{\text{cov}(x,y)}{\sigma_x . \sigma_y}$

3) $\text{cov}(x,y)$

$$= \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n}\sum (x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\,\bar{y})$$

$$= \frac{1}{n}\sum (x_i y_i) - \bar{x}.\frac{\sum y_i}{n} - \bar{y}.\frac{\sum x_i}{n} + \frac{n.\bar{x}\,\bar{y}}{n}$$

$$= \frac{1}{n}\sum (x_i y_i) - \bar{x}\,\bar{y} - \bar{y}\bar{x} + \bar{x}\,\bar{y}$$

$$= \frac{1}{n}\sum (x_i y_i) - \bar{x}\,\bar{y}$$

4) After change of origin and scale variables x and y become

$\left(\dfrac{x_i - a}{b}\right)$ and $\left(\dfrac{y_i - c}{d}\right)$ [a, b, c, d are chosen arbitrarily]

Show that

$\gamma_{xy} = \gamma_{\left(\frac{x-a}{b}\right)\left(\frac{y-c}{d}\right)}$

5) Simply use the formula of Spearman's rank correlation coefficient.

$$\rho = 1 - \frac{6\sum D_i^2}{n(n^2 - 1)}$$

where D_i = absolute difference between the ranks of an individual.

$$= 1 - \frac{6.x20}{10(10^2 - 1)}$$

$$= 1 - \frac{\overset{2}{\cancel{6}} \times \overset{2}{\cancel{20}}}{\cancel{10} \times \underset{33}{\cancel{99}}}$$

$$= 1 - \frac{4}{33} = .88$$

Check Your Progress 3

1) Since the regression lines intersect at $(\bar{x}, \bar{y})$, $\bar{x}$ and $\bar{y}$ can be obtained by solving the two given equations,

$\bar{x} = 13$, $\bar{y} = 17$.

Start assuming, $8x - 10y + 66 = 0$ to be the x on y line

and $40x - 18y = 214$ to be the y on x line.

$$b_{xy} = \frac{5}{4} \text{ and } \quad b_{yx} = \frac{20}{9}$$

$$\therefore \gamma^2 = b_{xy}.b_{yx} = \frac{25}{9} > 1$$

which is impossible, therefore, our assumption was wrong.

$8x - 10y + 66 = 0$ is y on x line

and

$40x - 18y - 214 = 0$ is x on y line

$$\therefore b_{xy} = \frac{9}{20} \text{ and } b_{yx} = \frac{4}{5}$$

$$\therefore \gamma = \frac{9}{20} \times \frac{4}{5} = 0.6$$

(Why the positive square root is taken?)

Given, $\sigma_x^2 = 9$

$$b_{xy} = \gamma.\frac{\sigma_y}{\sigma_x} = \frac{4}{5}$$

using the above relation find σ_y.

$\sigma_y = 4$

2) $b_{sy} = \frac{1}{m}$, $b_{ys} = \gamma.\frac{\sigma_s}{\sigma_y}$ given

$$\frac{\sigma_s^2}{\sigma_y^2} = \frac{1}{4} \text{ and } \quad \gamma = 0.4$$

$$\therefore b_{sy} = 0.4 \times \frac{1}{2} = 0.2 = \frac{1}{m}$$

$\therefore m = 5$.

3) y on x line is given by $y - \bar{y} = b_{yx}(x - \bar{x})$

$b_{yx} = 0.94$

$\therefore$ y on x is given by $y = 0.94.x + 92.18$

when $x = 45$, $y = 134.5$.

Check Your Progress 4

1) $\gamma_{12.3} = +0.09$; $\gamma_{13.2} = +0.64$; $R_{1.23} = +0.71$

Check Your Progress 5

1) Hint:

$\text{Log } y = \log a + (\log b)\, x$

$y = A + Bx$

Apply method of OLS and normal equations are:

$$\sum y = An + B\sum x$$

$$\sum xy = A\sum x + B\sum x^2$$

Find A and B and then find a and b. Get the answer $y = 100(1.2)^x$

2) $y = 1.42 - 1.07x + 0.55x^2$.

15.13 EXERCISES

1) In order to find out the correlation coefficient between two variables x and y from 12 pairs of observations the following calculations were made.

$\sum x = 30$, $\sum y = 5$, $\sum x^2 = 670$, $\sum y^2 = 285$ and $\sum xy = 334$

On subsequent verification it was found that the pair (x, y) = (10, 14) was mistakenly copied as (x, y) = (11, 4). Find the correct correlation coefficient?

2) The regression equations involving variables are $Y = 5.6 + 1.2x$ and $X = 12.5 + 0.6Y$. Find the arithmetic means of x and y and the correlation coefficient between them?

3) Obtain the linear regression equation that you consider more relevant for the following bivariate data and give reasons why you consider it to be so?

Age	56	42	72	36	63	47	55	49	38	42	68	60
Blood Pressure	147	125	160	118	149	128	150	145	115	140	152	155

4) Explain the terms *explained variation and unexplained variation.* If the correlation of coefficient between the variables x and y is 0.92 what percentage of total variation remains unexplained by the regression equation?

5) In a music contest two judges ranked eight candidates in order of their performance as follows:

Individuals	A	B	C	D	E	F	G	H
First judge	5	2	8	1	4	6	3	7
Second judge	4	5	7	3	2	8	1	6

Find the rank correlation coefficient?

UNIT 16 PROBABILITY DISTRIBUTION

Structure

16.0 Objectives

16.1 Introduction

16.2 Elementary Concept of Random Variable

16.3 Probability Mass Function

16.4 Probability Density Function

16.5 Probability Distribution Functions

16.6 Moments and Moment Generating Functions

16.7 Three Important Probability Distributions

- 16.7.1 Binomial Distribution
- 16.7.2 Poisson Distribution
- 16.7.3 Normal Distribution

16.8 Let Us Sum Up

16.9 Key Words

16.10 Some Useful Books

16.11 Answer or Hints to Check Your Progress

16.12 Exercises

16. 0 OBJECTIVES

After going through this unit, you will be able to:

- understand the random variables and how they are inseparable to probability distributions;
- appreciate moment generating functions and their role in probability distribution; and
- solve the problems of probability, which fit into binomial, poisson and normal distributions.

16.1 INTRODUCTION

In the previous unit on probability theory, we discussed the deterministic and non-deterministic events and introduced to random variables, which are outcomes of non-deterministic experiments. Such variables are always generated with a particular pattern of probability attached to them. Thus, based on the pattern of probability for the different values of random variable, we can distinguish them. Once we know these probability distributions and their properties, and if any random variable fits in a probability distribution, it will be possible to answer any question regarding the variable. In this unit, we have defined the random variable and made a broad distinction of the probability distributions based on whether the random variable is continuous or not. Then we have discussed how the moments of a probability distribution describe the distribution completely; how the technique of moment generating function could be used to obtain the moments of any probability distribution. In the next

section, we discuss three most widely used probability distributions viz., binomial, poisson and normal.

16.2 ELEMENTARY CONCEPTS OF RANDOM VARIABLE

When a random experiment is performed, we are often not interested in the detailed results, but only in the value of some numerical quantity determined by the experiment. It is natural to refer to a quantity whose value is determined by the result of a random experiment as a random quantity.

A variable is called random when its occurrence depends on the chance factor. In other words, there is a probability that the variable will take a particular value. When we toss a coin, say thrice, the number of heads we obtain is a random variable. Suppose in this case, X denotes the number of heads. Clearly, X can take four values, 0, 1, 2, and 3. More importantly, there is a probability attached with every value of the variable X. Therefore, X is a random variable. We can cite several examples of random variable: The number of red cards drawn when we draw 10 cards from a pack of 52 cards, the number we obtain when a die is rolled, number of accidents in a city and the number of printing mistakes in a page of a news paper.

Random variables could be either continuous or discrete. A discrete random variable takes only discrete or distinct values. In the above examples all are discrete variables. If Y is a variable that denotes the number when we roll a die, it will always take integral values. Practically, it will take any value among {1, 2, 3, 4, 5, 6} and cannot take the value 5.5 or 4.3. Therefore, Y is discrete random variable.

Similarly, a random variable is continuous if it can take any value in its range. The life of an electrical gadget, duration of phone calls received by a telephone operator and the amount of annual rainfall in a particular district of India are examples of continuous random variables. A random variable always takes real values. Therefore, we can define random variable as a real valued function defined over the points of a sample space (set of all possible outcomes of an experiment) with a probability measure.

16.3 PROBABILITY MASS FUNCTION

Probability distribution of a random variable is a statement specifying the set of its possible values together with the respective probabilities. When a random experiment is theoretically assumed to serve as a model, the probabilities could be represented as a function of the random variable.

Let a discrete random variable X assumes the values $x_1, x_2, x_3 \ldots x_n$ with probabilities $p_1, p_2, p_3, \ldots\ldots, p_n$, satisfying the condition $\sum_{i=1}^{n} p_i = 1$ The specification of the values of x_i together with the probabilities p_i defines the probability distribution of the random variable X. It is called discrete probability distribution of X.

Often the probability of the discrete random variable X assumes the value x is represented by f (x), f(x) = P (X = x) = probability that X assumes value x. The function f(x) is known as the probability mass function (p.m.f.).

The discrete random variable X can assume countably infinite number of values with p.m.f. f(x). The p.m.f. f(x) must satisfy following two conditions:

1) $f(x) \geq 0$

2) $\sum_{i=1}^{n} f(x_i) = 1$

Example: An unbiased coin is tossed until the first head is obtained. If the random variable X denotes the number of tails preceding the first head, then what is the probability distribution of X?

Values of X	f(X=x)
0	½
1	$(½)^2$
2	$(½)^3$
3	$(½)^4$
4	$(½)^5$
....	.....
....	.....
N	$(½)^n$

The probability distribution of the random variable X is shown in the table. Clearly, the p.m.f. of X is $f(X=x) = (1/2)^n$. It satisfies the prerequisites mentioned earlier, i.e., f(x) is always greater than '0' and $\sum_{i=1}^{n} f(x) = 1$. Note that X is a countably infinite random variable.

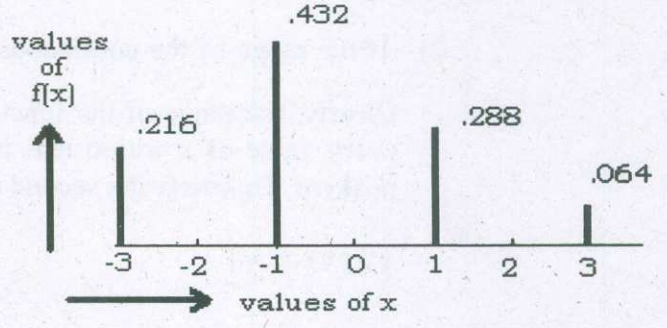
$\sum_{i=1}^{n} f(x) = 1$

$= ½ + (½)^2 + (½)^3 + (½)^4 + \ldots\ldots + (½)^n = ½\{1 + (½) + (½)^2 + (½)^3 + \ldots\ldots + (½)^{n-1}\}$

$= ½ . 1/(1 - ½) = 1.$

Example: Let X be a discrete random variable. Its probability distribution is given by the following table. Its density could be represented by the accompanying figure.

Values of X	f(X=x)
-3	0.216
-1	0.432
1	0.288
3	0.064



16.4 PROBABILITY DENSITY FUNCTION

In the previous section, we discussed the probability distribution of a discrete random variable. But what if the random variable is continuous, i.e., it can take any value in its range. Since the number of possible values a continuous variable can take is uncountable infinite, we cannot assign a probability to each value taken by the variable, as we did in case of the discrete random variable. In case of continuous random variable, we assign probability to an interval, which is in the range of the relevant random variable. A continuous random variable on the other hand has distribution f(x) is a continuous non-negative function, which gives the probability that the random variable will lie in a specified interval when integrated over the interval, i.e.,

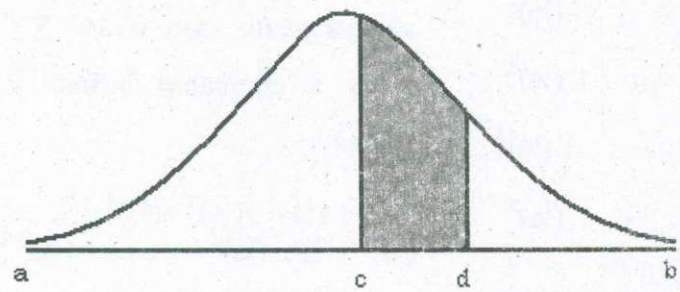
$P(c \leq x \leq d) = \int_{c}^{d} f(x)$

The function f(x) is called the probability density function (p.d.f.) provided it satisfies following two conditions

1) $f(x) \geq 0$

2) If the range of the continuous random variable is (a, b) $\int_a^b f(x) = 1$

The shaded area in the following figure represents the probability that the variable x will take values between the interval (c, d), whereas its range is (a, b). In the figure, we have taken the values of x in the horizontal axis and those of f(x) on the vertical axis. This is known as the probability curve. Since f(x) is a p.d.f., total area under the probability curve is 1 and the curve cannot lie below the horizontal axis as f(x) cannot take negative values. Any function that satisfies the above two conditions, can serve as a probability density function in the presence of real values only.



Example: If x is a continuous random variable with the probability density function $f(x) = k.e^{-3x}$ for $x > 0$ and 0 otherwise, then find k and P $(0.5 \leq x \leq 1)$

If the function specified in the problem is to serve as a p.d.f., then the following two conditions must hold:

1) $f(x) \geq 0$ for all values of x

2) If the range of the continuous random variable is (a, b), then $\int_a^b f(x) = 1$.

Clearly, the range of the function given in the problem is $-\infty$ to ∞ and for every value of x within that interval, f(x) is positive, provided that 'k' is positive. To satisfy the second condition

$$\int_0^\infty f(x)\,dx = 1$$

$$or, \int_0^\infty f(x)dx + \int_\infty^0 f(x)dx = 1$$

$$or, \int_0^\infty f(x)\,dx = 1$$

$$or, \int_0^\infty ke^{-3x}\,dx = 1$$

or, $k/3 = 1$

or, $k = 3$

Thus, we get, $f(x) = 3.e^{-3x}$ for $x > 0$ and 0 otherwise.

$$P(0.5 \leq x \leq 1) = \int_{0.5}^{1} 3e^{-3x}\,dx = -e^{-3} + e^{-1.5} = 0.173$$

16.5 PROBABILITY DISTRIBUTION FUNCTIONS

If X is a *discrete random variable* and the value of it's probability at the point t is given by f(t), then the function given by

$F(x) = \sum_{x \le t} f(t)$ for $-\infty \le x \le \infty$

It is called the distribution function or the cumulative distribution of X and is given by the summation of the probabilities if the random variable X takes values less than x.

The density function of a discrete random variable satisfies following conditions:

i) $F(-\infty) = 0, F(\infty) = 1$

ii) If $a < b$, then $F(a) \le F(b)$ where a and b are any real number.

If X is a *continuous random variable* and the value of it's probability density at the point t is given by f(t), then the function given by

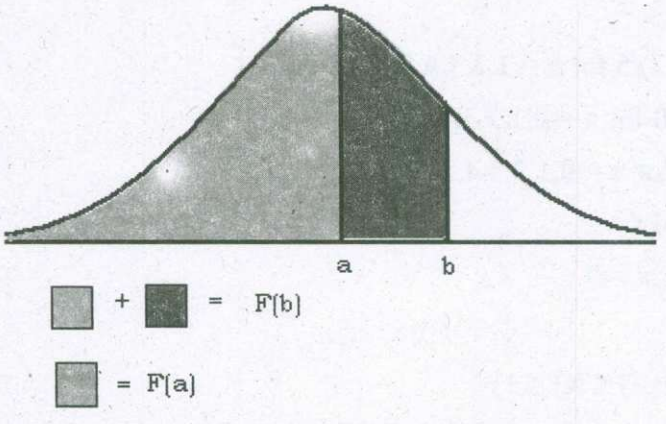
$F(x) = P(X \le x) = \int_{\infty}^{x} f(t)dt$ is called the distribution function. The distribution function of a continuous random variable has the same nice properties as that of a discrete random variable viz.,

i) $F(-\infty) = 0, F(\infty) = 1$

ii) If $a < b$ then $F(a)) \le F(b)$ where a and b are any real number

iii) Furthermore, it follows directly from the definition that

$P(a \le x \le b) = F(b) - F(a)$, where a and b are real constants with $a \le b$.



iv) $f(x) = \frac{d}{dx}F(x)$ where the derivatives exist.

Example: Find the distribution function of the random variable x of the previous example and use it to reevaluate $P(0.5 \le x \le 1)$

For all non-positive values of x, f(x) takes the value 0. Therefore,

$F(x) = 0$ for $x \le 0$.

For $x > 0$,

$F(x) = \int_{\infty}^{x} f(t)dt$

or $\int_0^x 3e^{-3t}\,dt$

or, $1 - e^{-3x}$

Thus, $F(x) = 0$ for $x \leq 0$, and $F(x) = 1 - e^{-3x}$ for $x > 0$.

To determine $P(0.5 \leq x \leq 1)$, we use the fourth property of the distribution function of a continuous random variable.

$P(0.5) \leq x) \leq 1) = F(1) - F(0.5) = (1 - e^{-3)} - (1 - e^{-3 \times 0.5}) = 0.173$

Check Your Progress 1

1) What is the difference between probability mass function and probability density function? What are the properties that p.d.f. or p.m.f. must satisfy?

2) If X is a discrete random variable having the following p.m.f.,

X	P (X=x)
0	0
1	k
2	2k
3	2k
4	3k
5	k^2
6	$2k^2$
7	$7k^2 + k$

(i) determine the value of the constant k.

(ii) find P (X<5)

(iii) find P (X>5)

3) For each of the following, determine whether the given function can serve as a p.m.f.

i) $f(x) = (x - 2)/5$ for $x = 1,2,3,4,5$

ii) $f(x) = x^2/30$ for $x = 0,1,2,3,4$

iii) $f(x) = 1/5$ for $x = 0,1,2,3,4,5$

4) If X has the p.d.f.

$f(x) = k.e^{-3x}$ for $x > 0$

0 otherwise.

Find k and $P(0.5) \leq X) \leq 1)$

5) Find the distribution function of for the above p.d.f. and use it to reevaluate

$P(0.5) \leq X \leq 1)$

16.6 MOMENTS AND MOMENT GENERATING FUNCTIONS

Moments

If X is a discrete random variable, which takes the values $x_1, x_2, x_3 \ldots\ldots, x_n$ and f(x) is the probability that X will take the value x, the expected value of X is given by $E(X) = \sum_{i=1}^{n} x_i f(x_i)$

Correspondingly, if X is a continuous random variable and f(x) gives the probability density at x, the expected value of x is given by E (X) = $\int_{-\infty}^{\infty} x\, f(x) dx$

In the definition of continuous random variable, of course we assume that the integral exists, otherwise, the mathematical expectation is undefined. The expected value of a random variable gives its mean value if we think f(x) as the relative frequency of the random variable X when it takes the value x.

Again, if g (x) is a continuous function of the value of the random variable X, then expected value of g (x) is given by E (g (x)) = $\sum_{x=1}^{n} g(x_i) f(x_i)$, when X is a discrete random variable. When it is a continuous random variable, the expected value of g (x) is given by E (g (x)) = $\int_{-\infty}^{\infty} g(x) f(x) dx$.

The determination of mathematical expectation can often be simplified by using the following theorems (proofs of the theorems are discussed in the unit on probability)

1) E (a + b. X) = a + b. E (X)

2) E (a) = a, where a is a constant.

3) If $c_1, c_2, c_3, \ldots\ldots$ c_n are constants, $E\left[\sum_{i=1}^{n} c_i g(x_i)\right] = \sum_{i=1}^{n} c_i E\left[g(x_i)\right]$

In statistics as well as economics, the notion of mathematical expectation is very important. It is a special kind of moment. We will introduce the concept of moments and moment generating functions in the following.

The r^{th} order moment about the origin of a random variable x is denoted by μ'_r and given by the expected value of x^r.

Symbolically, for discrete random variable $\mu'_r = \sum_{i=1}^{n} x_i^r f(x_i)$, for r = 1, 2, 3... n.

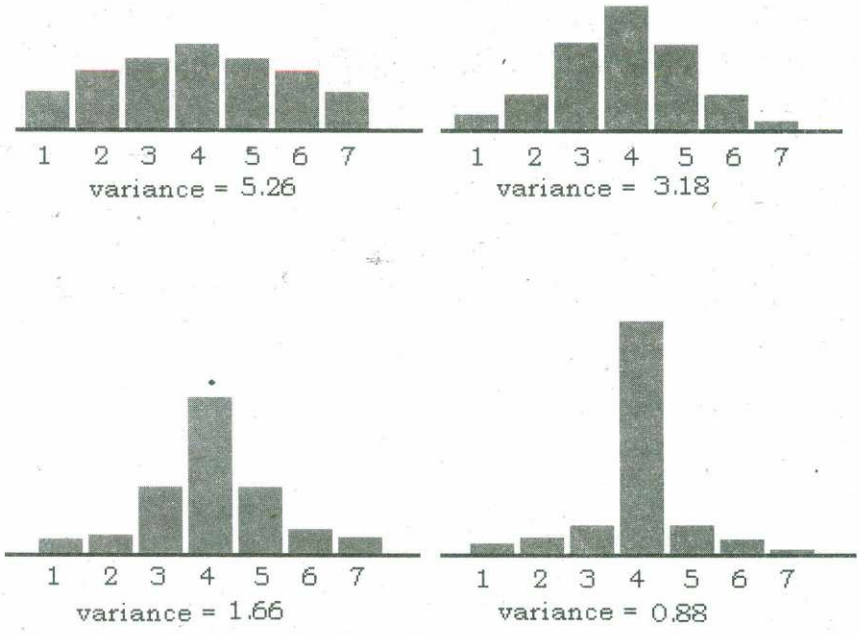
For continuous random variable $\mu'_r = \int_{-\infty}^{\infty} x^r f(x) dx$. It is interesting to note that the term moment comes from physics. If f(x) symbolizes quantities of points of masses, where x is discrete, acting perpendicularly on the x axis at distance x from the origin, then μ'_1 as defined earlier would give the center of gravity, which is the first moment about the origin. Similarly, μ'_2 gives the moments of inertia.

In statistics, μ'_1 gives the mean of a random variable and it is generally denoted by μ.

The special moments we shall define are of importance in statistics because they are useful in defining the shape of the distribution of a random variable, viz.,the shape of it's probability distribution or it's probability density.

The r^{th} moment of random variable about mean is given by μ_r. It is the expected value of $(X - \mu)^r$, symbolically, $\mu_r = E((X - \mu)^r) = \sum_{i=1}^{n} (x_i - \mu)^r f(x_i)$, for r = 1, 2, 3, ..., n, for discrete random variable and for continuous random variable it is given by $\mu_r = \int_{-\infty}^{\infty} (x - \mu)^r . f(x)\, d\, x$.

The second moment about mean is of special importance in statistics because it gives an idea about the spread of the probability distribution of a random variable. Therefore, it is given a special symbol and a special name. The variation of a random variable and it's positive square root is called standard deviation. The variance of a variable is denoted by Var (X) or V (X) or simply by σ^2.



Distributions with same mean but different variances

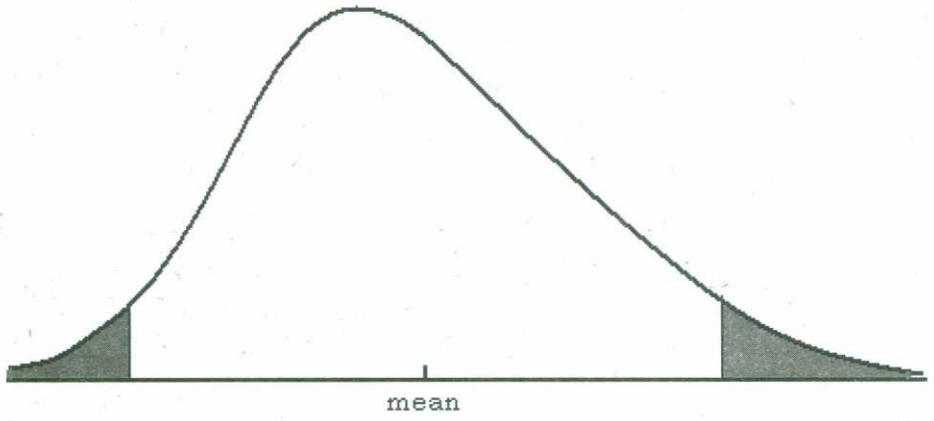
The above example shows how probability distributions vary with the variance of the random variable. A high value of σ^2 means the spread of the distribution is thick at the tails and a low value of σ^2 implies the spread of the distribution is tall at the mean of the distribution and the tails are flat. Similarly, third order moment about mean describes the symmetry or skewness (i.e., lack of symmetry) of the distribution of a random variable.

We state few important theorems on moments without going into details as they have been covered in the unit on probability.

1) $\sigma^2 = \mu'_2 - \mu^2$
2) If the random variable X has the variance, then Var $(a.X + b) = a^2 \sigma^2$
3) **Chebyshev's theorem**: To determine how σ or σ^2 is indicative of the spread or the dispersion of the distribution of the random variable, Chebyshev's theorem is very useful. Here we will only state the theorem:

If μ and σ^2 are the mean and variance of a random variable, say X, then for any constant k the probability is at least ($1 - 1/k^2$) that X will take on a value within k standard deviations (k. σ); symbolically

$$P(|X - \mu| < k.\sigma) \geq 1 - 1/k^2$$



Moment Generating Functions

Although moments of most distributions can be determined directly by evaluating the necessary integrals or sums, there is an alternative procedure, which sometimes provides considerable simplifications. This technique is known as the technique of moment generating function.

The moment generating function of a random variable X, where it exists is given by,

$M_x(t) = E(e^{tX}) = \sum_{i=1}^{n} e^{tx} f(x) =$ when X is discrete, and

$M_x(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx$ when X is continuous

To explain why, we refer to this function as a "moment generating function". Let us substitute for e^{tx} with its Maclaurin's series expansion as follows:

$e^{tx} = 1 + tx + t^2x^2/2! + t^3x^3/3! + \ldots\ldots + t^rx^r/r! + \ldots\ldots\ldots\ldots$

Thus, for discrete case we get,

$$M_x(t) = E(e^{tx}) = \sum_{i=1}^{n} e_i\, f(x) = \sum_{i=1}^{n} (1 + tx + t^2x^2/2! + t^3x^3/3! + \ldots\ldots + t^rx^r/r!$$

$$+\ldots\ldots).f(x) = \sum_{i=1}^{n} f(x) + t.\sum_{i=1}^{n} x.f(x) + t^2/2!.\sum_{i=1}^{n} x^2f(x) + t^3/3!.\sum_{i=1}^{n} x^3f(x)$$

$$+\ldots\ldots + t^r/r!.\sum_{i=1}^{n} x^rf(x) + \ldots\ldots\ldots\ldots = 1 + t.\mu + t^2/2!.\mu'_2 + \ldots\ldots + t^r/r!.\mu'_r$$

$+\ldots\ldots\ldots\ldots$

Thus, we can see that in the Maclaurin's series of the moment generating function of X, the coefficient of $t^r/r!$ is μ'_r, which is nothing but the r^{th} order moment about the origin. In the continuous case, the argument is the same (readers many verify that).

To get the r^{th} order moment about the origin, we differentiate $M_x(t)$ r times with respect to t and put t = 0 in the expression obtained. Symbolically,

$\mu'_r = d^r M_x(t)/dt^r|_{t=0}$

An example will make the above clear.

Example: Find the moment generating function of the random variable whose probability density is given by

$F(x) = e^{-x}$ if $x > 0$

0 otherwise

and use that to find the expression for μ'_r.

By definition,

$$M_x(t) = E(e^{tx}) = \int_{-\infty}^{0} et^r f(x)dx = \int_{-\infty}^{\infty} e^{-r}\ (1-t)\,dx = 1/1-t \text{ for } t<1.$$

When $|t|<1$, the Maclaurin's series for this moment generating function is

$M_x(t) = 1 + t + t^2 + t^3 + t^4 + \ldots\ldots + t^r + \ldots\ldots = 1 + 1!.t/1! + 2!.t^2/2! + 3!.t^3/3! + + \ldots\ldots + r!.t^r/r! + \ldots\ldots$

Hence, $\mu'_r = d^r M_x(t)/dt^r|_{t=0} = r!$ for $r = 0, 1, 2\ldots$

If 'a' and 'b' are constants, then

1) $M_{x+a}(t) = E(e^{t(x+a)}) = e^{at}.M_x(t)$

2) $M_{x.b}(t) = E(e^{t.b.x}) = M_x(b.t)$

3) $M_{(x+a)/b}(t) = E(e^{t(x+a)/b}) = e^{(a/b)t}.M_x(t/b)$

Among the above three results, the third one is of special importance. When $a = -\mu$ and $b = \sigma$, $M_{(x-\mu)/\sigma}(t) = E(e^{t(x-\mu)/\sigma}) = e^{(-\mu/\sigma)t}.M_x(t/\sigma)$

Check Your Progress 2

1) Given X has the probability distribution $f(x) = 1/8.\ {}^3C_x$ for $x = 0, 1, 2$, and 3, find the moment generating function of this random variable and use it to determine μ'_1 and μ'_2.

.....

.....

.....

.....

.....

16.7.1 Binomial Distribution

Repeated trials play has a very important role in probability and statistics, especially when the number of trials is fixed and the probability of success (or failure) in each of the trial is independent.

The theory, which we discuss in this section, has many applications; for example, it applies to events like the probability of getting 5 heads in 12 flips of a coin or the probability that 3 persons out of 10 having a tropical disease will recover. To apply binomial distribution in these cases, the probability (of getting head in each flip and recovering from the tropical disease for each person) should exactly be the same. More importantly, each of the coin tossed and chance of recovering of each patient should be independent.

To derive a formula for the probability of getting 'x successes in n trials' under the stated conditions we proceed as follows: suppose the probability of getting a success is 'p' since every experiment has only two possible outcomes (this type of events are called Bernoulli events) - the probability of a success or a failure. The probability of failure is simply '1 – p' and the trials or the experiments are independent of each other. The probability of getting 'x' successes and 'n – x' failures in n trials is given by $p^x (1-p)^{n-x}$. The probabilities of success and failure are multiplied by virtue of the assumption

that these experiments are independent. Since the probability applies to any sequence of n trials in which there are 'x' successes and 'n – x' failures, we have to count how many sequences of this kind are possible and then we have to multiply, $p^x.(1-p)^{n-x}$, by that number. Clearly, the number of ways in which we have 'x' successes and 'n – x' failures is given by ${}^nC_x$. Therefore, the desired probability of getting 'x' success in 'n' trials is given by ${}^nC_x p^x(1-p)^{n-x}$. Remember that the binomial distribution is a discrete probability distribution with parameters n and p.

A random variable X is said to have a binomial distribution and is referred as a binomial random variable, if and only if it's probability distribution is given by

$B(x;n,p) = {}^nC_x.\ p^x.(1-p)^{n-x}$ for x=0, 1, 2, ……..,n

Example: Find the probability of getting 7 heads and 5 tails in 12 tosses of an unbiased coin.

Substituting x = 7, n = 12, p = ½ in the formula of binomial distribution, we get the desired probability of getting 7 heads and 5 tails.

$B(7;12,1/2) = {}^{12}C_7.\ 1/2^{7}.(1-1/2)^{5} = {}^{12}C_7.\ (1/2)^{12}$

There are a few important properties of the binomial distribution. While discussing these, we retain the notations used earlier in this unit.

- The mean of a random variable, say X, which follows binomial distribution is $\mu = n.\ p$ and variance $\sigma^2 = n.\ p.(1-p) = n.\ p.\ q$ (the proofs of these properties are left as recap exercises)
- $b(x; n, p) = b(n - x, n, 1- p)$

  $b(n - x, n, 1- p) = {}^nC_{n-x}\ (1-p)^{n-x}.p^x = b(x; n, p)$ [as ${}^nC_{n-x} = {}^nC_x$]
- Binomial distribution may have either one or two modes. When (n+1)p is not an integer, mode is the largest integer contained therein. However, when (n+1)p is an integer, there are two modes given by (n + 1)p and {(n + 1)p -1}.
- Skewness of binomial distribution is given by $(q - p)/\sqrt{n.p.q}$, where q = (1-p)
- The kurtosis of binomial distribution is given by (1– 6pq)/npq
- If X and Y are two random variables, where X follows binomial distribution with parameters (n_1, p) and Y follows binomial distribution with parameters (n_2, p), then the random variable (X +Y) also follows binomial distribution with parameters ($n_1 + n_2$, p).
- Binomial distribution can be obtained as a limiting case of hypergeometric distribution.
- The moment generating function of a binomial distribution is given by

$$M_x(t) = E(e^{tx}) = \sum_{i=0}^{n} e^{tx} f(x) = \sum_{i=0}^{n} e^{tx}\ .\ {}^nC_x\ p^x\ (1-p)^{n-x}$$

$= \sum_{i=0}^{n} {}^nC_x\ (pe^t)^x\ (1-p)^{n-x}$. The summation is easily recognized as the binomial expansion of $\{p.e^t + (1-p)\}^n = \{1 + p(e^t - 1)\}^n$

We can derive the mean and variance of the binomial distribution using the moment generating function.

If we differentiate $M_x(t)$ with respect to t twice, we get

$M'_x(t) = npe^t[1 + p(e^t - 1)]^{n-1}$

$M''_x(t) = npe^t(1 - p + npe^t)[1 + p(e^t - 1)]^{n-2}$

And substituting $t = 0$, we get $\mu'_1 = np$ and $\mu'_2 = np(1 - p + np)$

Thus, $\mu = n.p$ and $\sigma^2 = \mu'_2 - (\mu'_1)^2 = np(1 - p + np) - (np)^2 = n.p.(1-p)$

16.7.2 Poisson Distribution

When the number of trials discussed in the case of binomial distribution is very large, the calculation of the probabilities with binomial distribution becomes very cumbersome. Suppose we want to know what is the probability that in a clinical test 200 out of 300 mice will survive, after being infected by a virus. There is 50% chance that each mouse can fight with the virus by producing antibodies. If we use binomial distribution, the required probability is given by $^{300}C_{200}$. ½ 300. In this section, we will discuss a probability distribution, which can be used to approximate binomial probabilities of this kind. Specifically, we will investigate the limiting form of the binomial distribution when $n \to \infty$ and $p \to 0$, while n.p remains constant. Let us suppose $\lambda = n.p$, therefore $p = \lambda / n$. We can also write the binomial distribution as

$b(x,n,p) = {}^nC_x (\lambda / n)^x (1 - (\lambda / n))^{n-x} = n(n-1)(n-2)(n-3).....(n-x+1)/x! \times (\lambda / n)^x (1 - (\lambda / n))^{n-x} = 1.(1-1/n)(1-2/n)(1-3/n)...(1-(x-1)/n)/x! \times \lambda^x \times [(1-(\lambda/n))^{-n/\lambda}]^{-\lambda} \times (1-(\lambda/n))^{-x}$

Finally, if we have $n \to \infty$ while x and p are constants

$(1-1/n)(1-2/n)(1-3/n)...(1-(x-1)/n) \to 1$

$(1-(\lambda/n))^{-x} \to 1$

$(1-(\lambda/n))^{-n/\lambda} \to e$

Therefore, the limiting form of the binomial distribution becomes

$P(x, \lambda) = \lambda^x e^{-\lambda}/x!$ for $x = 0, 1, 2, 3, ...$

Thus, in the limit when $n \to \infty$, $p \to 0$ and $np = \lambda$ remains constant; the number of success is a random variable and will follow a poisson distribution with only parameter λ. This distribution is named after the French mathematician Simeon Poisson. Following random variables are classic examples of poisson distribution.

1) Number of accidents on a road crossing.
2) Number of defects per unit area of a sheet material.
3) Number of telephone calls received by a telephone attendant.
4) Number of suicides in a year in an area etc.

There are few important properties of the poisson distribution.

- Poisson distribution is a discrete probability distribution, where the random variable assumes countably infinite number of values such as 0,1,2,3, ... to ∞. The distribution is completely specified if the parameter λ is known.
- Mean and variance of poisson distribution are the same, both being λ.
- Poisson distribution like the binomial distribution may have either one or two modes. When λ is not an integer, mode is the largest value contained in λ and when λ is an integer, there are two modes, λ and $(\lambda - 1)$.

- The distribution has skewness = $1/\sqrt{\lambda}$ and kurtosis = $1/\lambda$, therefore we can conclude that poisson distribution is positively skewed and leptokurtic.
- If X_1 and X_2 are two random variables following the poisson distribution with parameters λ_1 and λ_2 respectively, then the random variable $(X_1 + X_2)$ also follows poisson distribution with the parameter $(\lambda_1 + \lambda_2)$.
- As we have discussed earlier, the poisson distribution could be used as an approximation to binomial distribution when n is large but np is fixed.
- The moment generating function of the poisson distribution is given by $M_x(t) = E(e^{tX})$

$$= \sum_{i=0}^{\infty} e^{tx}.f(x)$$

$$= \sum_{i=0}^{\infty} e^{tx}.\lambda^x e^{-\lambda}/x!$$

$$= e^{-\lambda} \sum_{i=0}^{\infty} e^{tx}.\lambda^x/x!$$

$$= e^{-\lambda} \sum_{i=0}^{\infty} (\lambda e^t)^x/x!$$

In the above expression $\sum_{i=0}^{\infty} (\lambda e^t)^x/x!$ can be recognized as the Maclaurin's series of e^z where $z = \lambda e^t$.

Thus, the moment generating function of the poisson distribution is

$$M_x(t) = e^{-\lambda}.e^{\lambda et}$$

Differentiating $M_x(t)$ twice with respect to t we get,

$$M_x'(t) = \lambda.e^t.e^{\lambda(et-1)}$$

$$M_x''(t) = \lambda.e^t.e^{\lambda(et-1)} + \lambda^2.e^{2t}.e^{\lambda(et-1)}$$

Therefore, $\mu'_1 = M_x'(0) = \lambda$ and $\mu'_2 = M_x''(0) = \lambda + \lambda^2$ and we get $\mu = \lambda$ and $\sigma^2 = \mu'_2 - (\mu'_1)^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$

Example: Let X be a random variable following poisson distribution. If P(X = 1) =P(X = 2), find P (X = 0 or 1) and E (X).

For poisson distribution, the p.m.f. is given by $P(x, \lambda) = \lambda^x e^{-\lambda}/x!$

Therefore, $P(X = 1) = \lambda^1 e^{-\lambda}/1! = \lambda e^{-\lambda}$

$P(X = 2) = \lambda^2 e^{-\lambda}/2! = \lambda^2 e^{-\lambda}/2$

As P (X = 1) =P (X = 2), from the equation $\lambda e^{-\lambda} = \lambda^2 e^{-\lambda}/2$, we get $\lambda = 2$.

Therefore, $E(X) = \lambda = 2$ and

$P(X = 0 \text{ or } 1) = P(X = 0) + P(X = 1) = 2^0 e^{-2}/0! + 2^1 e^{-2}/1! = 3.e^{-2}$.

Example: In a textile mill, on an average, there are 5 defects per 10 square feet of cloth produced. If we assume a poisson distribution, what is the probability that a 15 square feet cloth will have at least 6 defects?

Let X be a random variable denoting the number of defects in a 15 square feet piece of cloth.

Since on an average, there are 5 defects per 10 square feet of cloth, there will be on an average 7.5 defects per 15 square feet of cloth i.e., $\lambda = 7.5$. We are to find $P(X \geq 6) = 1 - P(X \leq 5)$.

You are asked to verify the table for

X	P (X)
0	0.0006
1	0.0041
2	0.0156
3	0.0389
4	0.0729
5	0.1094

$\lambda = 7.5$ with the help of a calculator. From the table we obtain $P(X \leq 5) = .2415$. Therefore, $P(X \geq 6) = 1 - .2415 = .7585$.

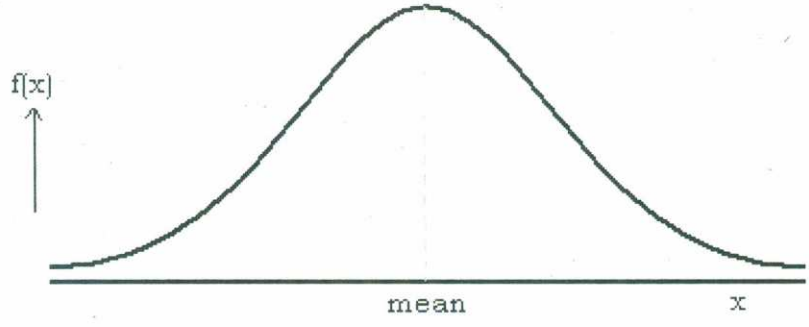
16.7.3 Normal Distribution

The normal distribution, which we will study in this section is the cornerstone of modern statistical theory. It was first investigated in the eighteenth century when scientists observed an astonishing degree of regularity in errors of measurement. They found that the patterns of the errors of measurement could be closely approximated by continuous curves, which they referred as the "normal curves". The mathematical properties of such continuous curves were first studied by Abraham de Moivre, Pierre Laplace and Karl Gauss.

A continuous random variable X is said to follow a normal distribution, and it is referred to as a normal random variable, if and only if the probability density is given by

$$F(x) = n(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\Pi}} \times e^{-\frac{1}{2}\{(x-\mu)/\sigma\}^2} \quad \text{for } -\infty < x < \infty, \text{ where } \sigma > 0$$

The shape of normal distribution is like a cross section of a bell and is shown as,



While defining the p.d.f. of normal distribution we have used the standard notations, where σ stands for standard deviation and μ for mean of the random variable X. Note that f(x) is positive as long as σ is positive which is guaranteed by the fact that standard deviation of a random variable is always positive. Since f(x) is a p.d.f. while X can assume any real value, integrating f(x) over $-\infty$ to ∞ we should get the value 1. In other words, the area under the curve must be equal to 1. Let us prove that

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\Pi}} \times e^{\frac{1}{2}\{(x-\mu)/\sigma\}^2} \, dx = 1$$

We substitute (x - μ)/σ = z in the R.H.S of the above equation to get

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\Pi}} \times e^{\frac{1}{2}\{(x-\mu)/\sigma\}^2} \, dx = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = \frac{2}{\sqrt{2\Pi}} \int_{0}^{\infty} e^{-\frac{1}{2}z^2} dz$$

[Since the p.d.f. is symmetrical , integrating the function from 0 to ∞ twice is the same as integrating the same function $-\infty$ to ∞]

$$= \frac{2}{\sqrt{2\Pi}} \times \Gamma\left(\frac{1}{2}\right)/\sqrt{2} \text{ [since } \int_{0}^{\infty} e^{-\frac{1}{2}z^2} dz = \Gamma\left(\frac{1}{2}\right)/\sqrt{2} \text{]}$$

$= 2/\sqrt{2\Pi} \times \sqrt{\Pi} / \sqrt{2}$

= 1 .....[proved]

Normal distribution has many nice properties, which make it amenable to be applied to many statistical as well as economic models.

Example: The height distribution of a group of 10,000 men is normal with mean height 64.5" and standard deviation 4.5. Find the number of men whose height is

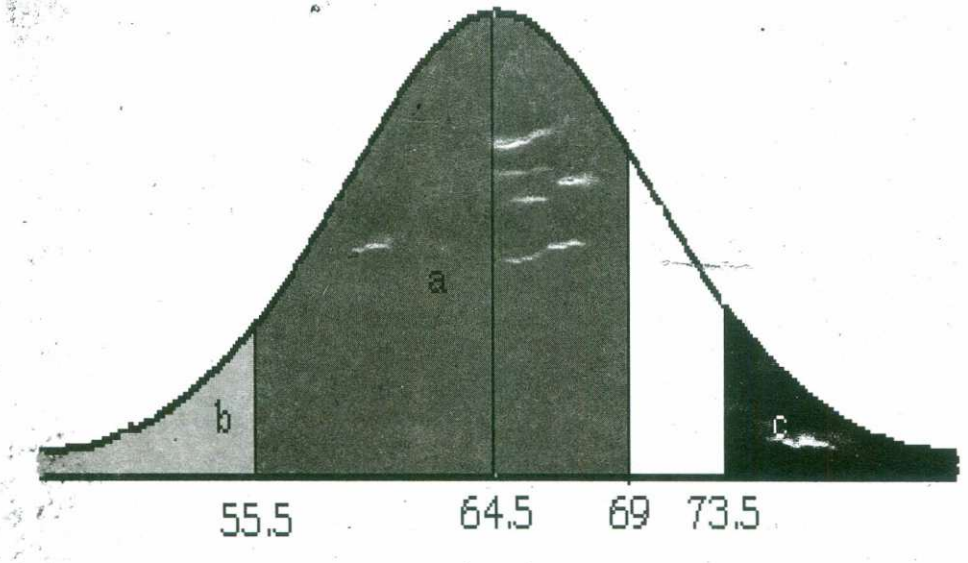
a) less than 69" but greater than 55.5"

b) less than 55.5"

c) more than 73.5"

The mean and standard deviation of a normal distribution is given by

μ = 64.5" and σ = 4.5. We explain the problem graphically. From the figure below, we can easily comprehend what we are asked to do. We are to find out the shaded regions but as we know area under a standard normal curve only, we have to reduce the given random variable into a standard normal variable



Let X is the continuous random variable measuring the height of each man.

Therefore, (X – 64.5)/4.5 = z is a standard normal variable.

The following table shows values of z for corresponding values of x. In the table true

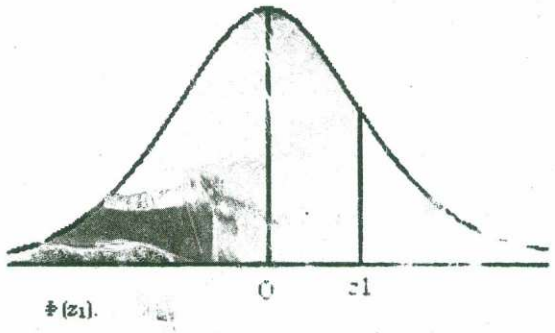
X	Z
55.5	-2
64.5	0
69	1
73.5	2

area under standard normal curve is given for only positive values of the standard normal variable and as the distribution is symmetrical, area under the curve for negative values of the standard normal variable is easy to find out. For standard normal curve (say for the variable z) the area under the curve to the left is conventionally denoted by $\Phi(z_1)$. This is shown in the figure below.

a) P (55.5<X<69) = P (-2<z<1) = Φ (1) - Φ (-2) = .82, Therefore, men of height less than 69" but greater than 55.5" is 10000×0.82 = 8200

b) P (X<55.5) = P (z<-2) = .02, Therefore, men of height less than 55.5" is 10000×0.02 = 200

c) P (X>73.5) = P (z > 2) = 1 - P (z < 2) = 1 - .98 = .02, Therefore, men of height greater than 73.5" is 10000×0.02=200



Properties of Normal Distribution

- Normal distribution is a continuous probability distribution.
- Normal distribution has two parameters, namely, μ and σ.
- Mean and standard deviation of a normal distribution is given by μ and σ respectively.
- For a normal distribution, the mean, median and mode are the same, i.e. μ. As a corollary to the above property, we can say that the first and third quartiles are equidistant from the mean of the normal distribution. Approximately,

  $Q_1 = \mu - 0.67\times \sigma$ and

  $Q_3 = \mu + 0.67\times \sigma$
- All odd order central moments of the normal distribution are 0.

  In general $\mu_{2r} = 1.3.5.....(2r - 1)\sigma^{2r}$ for r = 1,2,3,.....

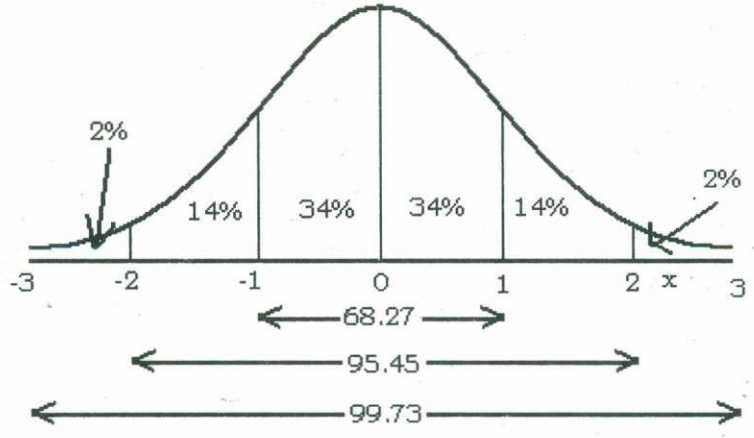
  $\mu_{2r+1} = 0$ for r = 1,2,3,.....
- The normal distribution is symmetrical as well as mesokurtic and

  skewness = 0 and kurtosis = 0.

- Normal distribution is symmetrical about it's mean. The two tails of the distribution are extended to infinity on both sides of the mean. The tails of the distribution never meet the horizontal axis. The maximum ordinate of the p.d.f. is at the mean, which is given by $1/\sigma\sqrt{2\Pi}$.
  - The points of inflextion of the normal curve are at $x = \mu + \sigma$ and $x = \mu - \sigma$ respectively. At these two points, normal curve changes its curvature.
  - If a random variable X follows normal distribution with the mean and variance μ and σ respectively, then the random variable $z = (x - \mu)/\sigma$ is called standard normal variable. It has a density function

$$f(x) = \frac{1}{\sqrt{2\Pi}} e^{\frac{-z^2}{2}} dz \quad \text{.....} -\infty < z < \infty$$

The continuous probability distribution defined above is known as standard normal distribution. In fact, this is a special kind of probability distribution with mean zero and standard deviation 1. The approximate area under the standard normal curve is shown in the following figure.



- If X and Y are two normal variables with mean μ_1 and μ_2 and standard deviation σ_1 and σ_2, then (X + Y) is also a normal variable with mean $(\mu_1 + \mu_2)$ and variance $(\sigma_1^2 + \sigma_2^2)$.
- The moment generating function of a normal curve is given by

$$M_X(t) = e^{\mu t + 1/2.\sigma^2 t^2}$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

The above expression could be written, after some algebraic manipulation as the following:

$$M_X(t) = e^{\mu t + \frac{1}{2}(t\sigma)2} \times \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-1/2\left[\frac{x-(\mu+t\sigma^2)}{\sigma}\right]} dx = e^{\mu t + \frac{1}{2}(t\sigma)2}$$

$$[\text{since } 1/\sigma\sqrt{2\Pi} \int_{-\infty}^{\infty} e^{-1/2\,[\{x-(\mu+t\sigma 2)\}/\sigma\,]2}\, dx = 1]$$

Differentiating $M_X(t)$ with respect to t twice, we can get

$$M'_X(t) = (\mu + \sigma^2 t)\, M_X(t)$$

$M''_X(t) = [(\mu + \sigma^2 t)^2 + \sigma^2]\, M_X(t)$

Substituting $t = 0$ in the above two equations

$M'_X(0) = \mu$

$M''_X(0) = \mu^2 + \sigma^2$. Therefore, E (X)= μ and Variance (X)= σ^2.

Check Your Progress 3

1) The mean and standard deviation of a binomial distribution is given by 4 and $\sqrt{\left(\frac{8}{3}\right)}$ respectively. Find the values of n and p.

2) Prove that poisson distribution is a limiting case of binomial distribution.

3) In turning out some toys in a manufacturing process in a factory, the average number of defectives is 10%. What is the probability of getting exactly 3 defectives in a sample of 10 toys chosen at random, by using poisson approximation to the binomial distribution? (take e = 2.72)

4) 2% of the items made by a machine are defective. Find the probability that 3 or more items are defective in a sample of 100 items? ($e^{-1} = 0.368$, $e^{-2} = .135$, $e^{-3} = .0498$)

5) The mean weight of 500 students at a University is 151 lbs and the s.d is 15. Assuming the weights are normally distributed, find how many students
i) weight between 120 and 155 lbs.

ii) more than 155 lbs.

Given: $\Phi(0.27) = 0.6064$; $\Phi(2.07) = 0.9808$; $\Phi(t)$ implies the area under the standard normal curve to the left of the ordinate at the point t.

6) The mean of a normal distribution is 50 and 5% of the values are greater than 60. Find the standard deviation of the distribution (given that the area under the standard normal curve between z = 0 and z = 1.64 is 0.45)

7) For a certain normal distribution, the first moment about 10 is 40, and the fourth moment about 50 is 48. Find the arithmetic mean and the standard deviation of the distribution.

8) Find the probability that 7 out of 10 persons will recover from a tropical disease if we can assume independence and the probability is 0.8 that any one of them will recover from the disease.

16.8 LET US SUM UP

In this unit we learnt the concepts like random variable, probability mass function and probability density function. You have been introduced to the three most elementary and most used distributions in theory of probability namely, Binomial, Poisson and Normal. First two of them are discrete and the last one being continuous. A distribution is defined mostly by its moments; therefore we have introduced the concepts concerning moments and how they are used to characterize a distribution. The moment generation function is a tricky tool of determining the moments of different distributions.

16.9 KEY WORDS

Binomial Distribution: A discrete probability distribution satisfying the following conditions is a binomial distribution:

1) It involves finite repetition of identical trials
2) Each trial has two possible outcomes: success and failure.
3) Trials are independent of each other.
4) The probability of the outcomes (success and failure) does not change from one trial to another.

Continuous Random Variable: If a random variable can take any value within its range then it is called a continuous random variable.

Discrete Random Variable: If a random variable takes only countable number of values and there is no possible values of the variable located between two juxtaposed values of the variable, then it is called a discrete random variable.

Normal Distribution: It is a continuous probability distribution with the following nice properties:

1) It is symmetrical about its mean;
2) All the measures of central tendency for this distribution is the same, i.e., for normal distribution Mean = Mode = Median; and
3) A variable following normal distribution can take any value within the range $(-\infty, \infty)$.

Probability Distribution: Probability distribution of a random variable is a statement specifying the set of its possible values together with the respective probabilities.

Probability Mass Function: Often the probability of the discrete random variable X assuming the value x is represented by f (x), f(x) = P (X = x) = probability that X assumes value x. The function f(x) is known as the probability mass function (p.m.f.) given that it satisfies following two conditions:

1) $f(x) \geq 0$
2) $\sum_{i=1}^{n} f(x_i) = 1$.

Probability Density Function: A continuous random variable having a probability distribution f(x) which is a continuous non-negative function, which gives the probability that the random variable will lie in a specified interval when integrated over the interval, i.e.,

$P(c \leq x \leq d) = \int_{c}^{d} f(x)$. The function f(x) is called the probability density function (p.d.f.) provided it satisfies following two conditions

1) $f(x) \geq 0$
2) the range of the continuous random variable is (a, b) $\int_{a}^{b} f(x) = 1$.

Poisson Distribution: A discrete probability distribution which is the limiting form of the binomial distribution, provided

1) The number of trials is very large in fact tending to infinity.
2) The probability of success in each trial is very small; tending to zero.

Random Variable: If a variable takes different values and for each of those values if there is a probability associated with it, the variable is called random variable.

16.10 SOME USEFUL BOOKS

Freund J.E. (2001), *Mathematical Statistics*, Prentice Hall of India.

Hoel, P (1962), *Introduction to Mathematical Statistics*, Wiley John & Sons, New York.

Hoel, Paul G. (1971), *Introduction to Probability Theory*, Universal Book Stall, New Delhi.

Olkin, I., L.J. Gleser, and C. Derman (1980), *Probability Models and Applications*, Macmillan Publishing, New York.

16.11 ANSWER OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) Do it yourself.

2) A p.m.f. must satisfy the following two properties

   1) $f(x) \geq 0$

   2) $\sum_{i=1}^{n} f(x_i) = 1.$

   Using the second property, we get $10k^2 + 9k - 1 = 0$. It gives two values of k, viz., -1 and 1/10. Clearly, k cannot take the value -1 (as $f(X = 1) = k$ and $f(x)$ is always non-negative). Given $k = 1/10$ rest is trivial algebra.

3) i) Cannot.

   ii) Cannot.

   iii) Cannot.

4) $f(x)$ to be a p.d.f., it must satisfy the condition $\int_{-\infty}^{\infty} f(x) = 1$, since in the problem $f(x)$ is zero for non positive values of x, the condition reduces to

   $$\int_{0}^{\infty} k.e^{-3x}\, dx = 1$$

   or, $k.\, [e^{-3x}/-3]_0^{\infty}$

or, k.1/3 = 1

or, k = 3

$$P(0.5 \leq X \leq 1) = \int_0^{\infty} 3.e^{-3x} = 0.173$$

5) $F(x) = \int_0^x f(t)dt = \int_0^x 3.e^{-3t}\, dt = 1 - e^{-3x}$

Since F(x) =0 for $x \leq 0$, we can write

$F(x) = 0$ for $x \leq 0$

$1 - e^{-3x}$ for $x > 0$

As we know $P(0.5 \leq X \leq 1) = F(1) - F(0.5) = 0.173$

Check Your Progress 2

1) $M_x(t) = E(e^{tX}) = \sum_{i=0}^{3} e^{tx} \frac{1}{8} {}^3C_x = 1/8\,(1 + 3e^t + 3e2^t + e3^t) = 1/8(1 + e^t)$

$\mu'_1 = M'_x(0) = 3/2$

$\mu'_2 = M'_x(0) = 3$

Check Your Progress 3

1) As the formula for mean and variance of binomial distribution are given by n.p and n.p. (1 - p) respectively, we get the following two equations

n.p = 4.....(1)

n.p. (1 - p) = 8/3.....(2)

Solving these two equations we get n = 12 and p = 1/3.

2) See Section 13.6.2.

3) $\lambda = 10 \times 0.1 = 1$, hence the probability of three defectives in the sample is given by

$f(3) = e^{-1} \times 1^3/3! = 0.061$

4) The number of defectives follows a binomial distribution. Since p= .02 and n=100 which is very large, making up, $n.p = \lambda = 100 \times .02 = 2$, a finite quantity, we use the poison approximation to binomial distribution.

P(of 3 or more defectives) = $1 - [f(0) + f(1) + f(2)] = 1 - e^{-2}\,[1 + 2 + 2^2/2!] = 0.325$

5) Clearly, the random variable denoting the weight, say X, of the students follow a normal distribution, mean (X)= 151;Var (X)= 15^2

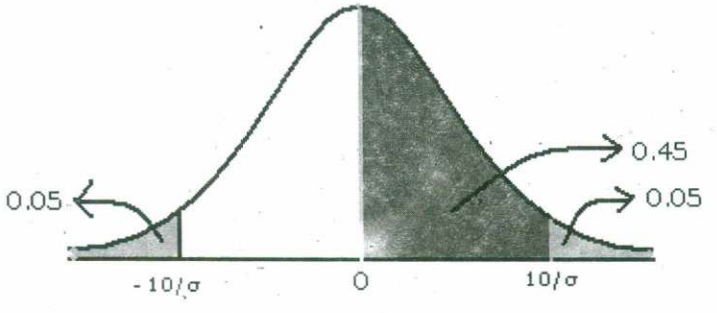
i) Proportion of students whose weight lie between 120 and 155 lbs = Area under the standard normal curve between the vertical lines at the standardized values; z = (120 - 151)/15 = -2.07 and z = (155 - 151)/15 = 0.27

$P(120 \leq X \leq 155) = \Phi(0.27) - \Phi(-2.07) = \Phi(0.27) - \{1 - \Phi(2.07)\}$
$= 0.6064 - 1 + 0.9808 = 0.5872$

ii) $P(X > 155) = 1 - P(X \leq 155) = 1 - \Phi(0.27) = 0.3936$

6) The probability that X takes values greater than 60 is 5% or 0.05. This must be the area under the standard normal curve to the right of the ordinate at the standardized value $z = (60 - 50)/\sigma = 10/\sigma$

Since the area to the right of z = 0 is 0.05 and the area between z = 0 and z = 1.64 is given to be 0.45, the area to the right of z = 1.64 is .5 - .45=.05. Thus, we get $10/\sigma = 1.64$, or, $\sigma = 6.1$.



7) Mean = A + First moment about A = 10 + 40 = 50.

Using the following formula

$\mu_{2r} = 1.3.5.....(2r - 1)\,\sigma^{2r}$ for r = 1,2,3,.....

Putting r = 2 in the above formula, we get,

$\mu_4 = 1.(4 - 1)\,\sigma^4 = 3\,\sigma^4$

since the fourth moment about 50 is 48, whereas the mean is 50.

Therefore, $3\,\sigma^4 = 48$

or, $\sigma = 2$.

8) Substituting x = 7, n = 10, and p = 0.8 into the formula for binomial distribution, we get

$b(7;10,0.8) = {}^{10}C_7.(0.8)^7(0.2)^3 = 0.2$ (approximately)

16.12 EXERCISES

1) Show that if a random variable has the following distribution

   $f(x) = \frac{1}{2}.e^{-|x|}$ for $-\infty\ x < \infty$

   Its moment generating function is given by $M_x(t) = 1/1 - t^2$.

2) Find the mean and the standard deviation of the random variable with the moment generating function $M_x(t) = e^{4(e^t - 1)}$.

3) For each of the following find the value of 'c' so that the function can serve as a probability distribution.

   i) $f(x) = c.x$ for $x = 1,2,3,4,5$

   ii) $f(x) = c.{}^5C_x$ for $x = 0,1,2,3,4,5$

   iii) $f(x) = c.x^2$ for $x = 1,2,3,4,5,\ldots..k$

   iv) $f(x) = c(1/4)^x$ for $x = 1,2,3,4,5\ldots\ldots\ldots$

4) Find the probability distribution function for a random variable whose density function is given by the following

$$F(x) = \begin{cases} 0\ for x \le 0 \\ x for 0 < x < 1 \\ 1\ for x \ge 1 \end{cases}$$

   And plot the graph of the distribution function as well as the density function.

5) Find the distribution function of the random variable X whose probability density is given by the following

$$f(x) = \begin{cases} x / 2 for 0 < x \le 1 \\ 1 / 2 for 1 < x \le 2 \\ (3 - x) / 2 for 2 < x < 3 \\ 0 \text{ otherwise} \end{cases}$$

   Draw the graph of the distribution and the density function.

6) What is the probability of guessing correctly at least 6 of the 10 answers in a true false objective test?

7) Show that the binomial distribution is symmetrical when $p = \frac{1}{2}$.

8) The average number of defects per yard on a piece of cloth is 0.9. What is the probability that a one-yard piece chosen at random contains less than two defects? [$e^{0.9} = 2.46$]

9) If the 5% of the electric bulbs manufactured by a company are defective, use suitable distribution to find the probability that in a sample of 100 bulbs

   i) None is defective

   ii) 5 bulbs will be defective [$e^{-5} = .007$]

10) Show that the probability that the number of heads in 400 tosses of a fair coin lies between 180 and 220 is approximately $2.\Phi(2) - 1$, where $\Phi(x)$ denotes the standard normal distribution function.

11) In a normal distribution 8% of the observations are under 50 and 10% are over 60. Find the mean and the standard deviation of the distribution? [Given that $\int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$ = 0.08 or 0.10 according as x = 1.4 or 1.28]

UNIT 17 PROBABILITY THEORY

Structure

17.0 Objectives
17.1 Introduction
17.2 Deterministic and Non-deterministic Experiments
17.3 Some Important Terminology
17.4 Definitions of Probability
17.5 Theorems of Probability
  17.5.1 Theorem of Total Probability
    17.5.1.1 Deductions from Theorem of Total Probability
  17.5.2 Theorem of Compound Probability
    17.5.2.1 Deductions from Theorem of Total Probability
17.6 Conditional Probability and Concept of Independence
  17.6.1 Conditional Probability
  17.6.2 Concept of Independent Events
17.7 Bayes' Theorem and its Application
17.8 Mathematical Expectations
17.9 Let Us Sum Up
17.10 Key Words
17.11 Some Useful Books
17.12 Answer or Hints to Check Your Progress
17.13 Exercises

17.0 OBJECTIVES

After going through this unit, you will be able to:

- understand the underlying reasoning of taking decisions under uncertain situations; and
- deal with different probability problems in accordance with theoretical prescriptions.

17.1 INTRODUCTION

Probability theory is a branch of mathematics that is concerned with random (or chance) phenomenon. It originated in the games related to chance and an Italian mathematician Jerome Cardan was first to write on the subject. However, the basic mathematical and formal foundation in the subject was provided by Pascal and Fermat. Contribution of Russian as well as European mathematicians helped the subject to grow.

17.2 DETERMINISTIC AND NON-DETERMINISTIC EXPERIMENTS

The agreement among scientists regarding the validity of most scientific theories rests, to a considerable extent, on the fact that the experiment on which the theory is based will yield the same result when they are repeated. If the same results are obtained when the experiment is repeated under the same conditions, we can conclude that the results are determined by the conditions,

or the experiment is deterministic. For example, any where in the world if we throw a stone to the sky it will certainly come back to the earth after some time.

However, there are experiments, which do not yield the same result even if the experimental conditions are kept constant. Examples are throwing a dice or picking a card from a well-shuffled pack of cards or tossing a coin. These experiments could be thought as the ones with unpredictable results. If we are willing to stretch our idea of experiment, then there are many examples of this kind in our day-to-day life. For example, two people living in same conditions will die at different and unpredictable ages. In literature, these experiments or events are referred to as "random experiments" or "random events". Thus, we define "non-deterministic" or "random experiments" as those *experiments* (experiment is an act which can be repeated under some given conditions) whose outcomes are not predictable before hand. Probability and branches of statistics are developed specially to deal with this kind of random events.

Each of the following may be called a random experiment:

1) Tossing a coin (or several coins)
2) Throwing a die (or several dice)
3) Drawing cards from a pack
4) Studying the distribution of boys and girls in families of India having two children
5) Drawing balls from an urn (or urns) having a given number of different kind of balls in each.

The result of a non-deterministic event is not predictable before hand, for there may be a number of outcomes associated with it. The same outcome of a random experiment could be described in several ways. In our 2nd example of random experiment, i.e., throwing a die, the possible outcomes are (1,2,3,4,5,6). These outcomes could have been described as "odd number of points" or "even number of points".

The term event in probability theory is used to denote any phenomenon, which occurs as a result of a random experiment. Events can be 'elementary' as well as 'composite'. An 'elementary' event cannot be decomposed into simpler events whereas a 'composite' event is an aggregate of several elementary events.

17.3 SOME IMPORTANT TERMINOLOGY

Sample Space: All possible outcomes of a non-deterministic experiment constitute the sample space of that experiment and are generally denoted by 'S'.

For example, if one coin is tossed, either head (H) or tail (T) will appear. Therefore, H and T constitute the sample space and we can write S = {H, D}.

Similarly if two coins are tossed, S = {(HH), (HT), (TH), (TT)}

A subset of sample space, which might be favorable to the occurrence of a particular event, is called a sub-sample space. For example, in the previous example, S_1 = {(HH), (HT), (TH)} is a sub sample space and it is favorable to the event that at least one head appears if two coins are tossed.

By N(S), we mean the number of events in S.

If two cubic dice are thrown N(S) = 6 × 6 = 36 and the S is given by

S= {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)

(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)

....(6,5), (6,6)}

Events: When some of the elements of all the possible outcomes in the sample space of a random experiment satisfy a particular criterion, we call it an event. If three unbiased coins are tossed, the sample space is

S = {(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)} and following events can be picked up from it:

E_1 = {(HHH)} = Event of getting all heads

E_2 = {(HHH), (HHT), (HTH), (THH)} = Event of getting at least 2 heads

E_2 = {(HHT), (HTH), (THH)} = Event of getting exactly 2 heads

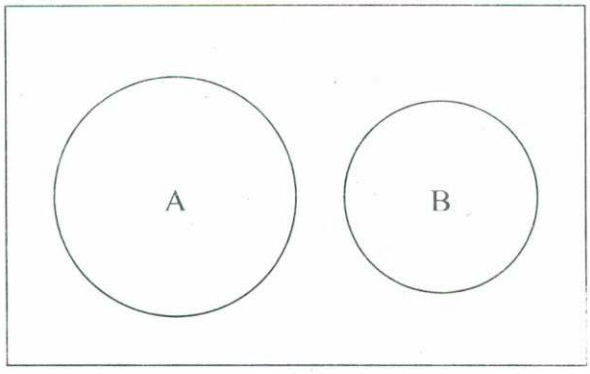
.....etc.

Mutually Exclusive Events: Events are said to be mutually exclusive if two or more of them cannot occur simultaneously. For example, while tossing a coin, the elementary events 'head' and 'tail' are mutually exclusive. Similarly, when we throw a die the appearance of the numbers 1, 2, 3,4,5,6, are mutually exclusive. We define the events from these experiments as follows:

A: the event of odd number of points; and

B: the event of even number of points.

These are also mutually exclusive.



In the above figure, the rectangular box represents the sample space and the circles represented by A and B represent sub sample spaces and contain favorable elements to the events A and B. The complete separation of the circles indicates that there is no element, which is common to both the events. Thus, A and B events are mutually exclusive. The above way of representing events is called Venn diagrams.

Mutually Exhaustive Events: Several events are said to be mutually exhaustive if and only if at least one of them necessarily occurs. For example, while tossing a coin, the events of head and tail are mutually exhaustive, as one of them must occur.

Equally Likely Events: The outcomes of a non-deterministic event are said to be equally Likely if occurrence of none of them can be expected in preference to another. For example, while tossing a coin, the occurrence of the event head or tail is equally likely if the coin is unbiased.

Independent Events: Events are said to be independent of each other if occurrence of one event is not affected by the occurrence of the others. For

example, while throwing a die repeatedly, the event of getting a '3' in the first throw is independent of getting a '6' in the second throw.

Conditional Events: When events are neither independent nor mutually exclusive, it is possible to think that one of them is dependent on the other. For example, it may or may not rain if the day is cloudy but if there is rain, there must be clouds in the sky. Thus, the event of rain is conditioned upon the event of clouds in the sky.

17.4 DEFINITIONS OF PROBABILITY

The term probability has been interpreted in terms of four definitions viz.,

1) Classical definition
2) Axiomatic definition
3) Empirical definition
4) Subjective definition

1) Classical Definition

The classical definition states that if an experiment consists of N outcomes which are mutually exclusive, exhaustive and equally likely and N_A of them are favorable to an event A, then the probability of the event A (P (A)) is defined as

$P(A) = N_A / N$

In other words, the probability of an event A equals the ratio of the number of outcomes N_A favorable to A to the total number of outcomes. See the following example for a better understanding of the concept.

Example1: Two unbiased dice are thrown simultaneously. Find the probability that the product of the points appearing on the dice is 18.

There are 36 (N) possible outcomes if two dice are thrown simultaneously. These outcomes are mutually exclusive, exhaustive and equally likely based on the assumption that the dice are unbiased. Now we denote A: the product of the points appearing on the dice is 18.

The events favorable to 'A' are [(3, 6), (6, 3)] only, therefore, $N_A = 2$.

According to classical definition of probability

$P(A) = N_A / N = 1/18$

When none of the outcome is favorable to the event A, $N_A = 0$, P (A) also takes the value 0, in that case we say that event A is impossible.

There are many defects of the classical definition of probability. Unless the outcomes of an event are mutually exclusive, exhaustive and equally likely, classical definition cannot be applied. Again, if the number of outcomes of an event is infinitely large, the definition fails. The phrase 'equally likely' appearing in the classical definition of probability means equally probable, thus the definition is circular in nature.

2) Axiomatic Definition

In the axiomatic definition of probability, we start with a probability space 'S' where set 'S' of abstract objects is called outcomes. The set S and its subsets

are called events. The probability of an outcome A is by definition a number P (A) assigned to A. Such a number satisfies the following axioms:

a) $P(A) \geq 0$ i.e., P (A) is nonnegative number.

b) The probability of the certain event S is 1, i.e., P (S) = 1.

c) If two events A and B have no common elements, or, A and B are mutually exclusive, the probability of the event (A U B) consisting of the outcomes that are in A or in B equals to sum of their probabilities:

$P(A \cup B) = P(A) + P(B)$

The axiomatic definition of probability is relatively recent concept (see Kolmogoroff, 1933). However, the axioms and the results stated above had been used earlier. Kolmogoroff's contribution was the interpretation of probability as an abstract concept and the development of the theory as a pure mathematical discipline.

We comment next on the connection between an abstract sample space and the underlying real experiment. The first step in model formation is between elements of S and experimental outcomes. The actual outcomes of a real experiment can involve a large number of observable characteristics. In the formation of the model, we select from these characteristics the one that is of interest in our investigation.

For example, consider the possible models of the throwing of an unbiased die by the 3 players X, Y and Z.

X says that the outcomes of this consist of six faces of the die, forming the sample space {1,2,3,4,5,6}.

Y argues that the experiment has only 2 outcomes, even or odd, forming the sample space {even, odd}

Z bets that the die will rest on the left side of the table and the face with one point will show. Her experiment consists of infinitely many points consisting of the six faces of the die and the coordinate of the table where the die rests finally.

3) Empirical Definition

In N trials of a random experiment if an event is found to occur m times, the relative frequency of the occurrence of the event is m/N. If this relative frequency approaches a limiting value p, as N increases indefinitely, then 'p' is called the probability of the event A.

$$P(A) = \lim_{N \to \infty} \left(\frac{m}{N} \right)$$

To give a meaning to the limit we must interpret the above formula as an assumption used to define P(A). This concept was introduced by Von Mises. However, the use of such a definition as a basis of deductive theory has not enjoyed wide acceptance.

4) Subjective Definition

In subjective interpretation of probability, the number P (A) is assigned to a statement. A, which is a measure of our state of knowledge or belief concerning the truth of A. These kinds of probabilities are most often used in our daily life and conversations. We often make statements like "I am 100% sure that I will pass the examination" i.e., P(of passing the examinations) = 1,

or "there is 50% chance that India will win the match against Pakistan" i.e., P(India will win the match against Pakistan)= ½

Check Your Progress 1

1) What is the probability that all three children born in a family will have different birthdays?

.....

.....

.....

.....

.....

2) Five persons a, b, c, d, e occupy seats in a row at random. What is the probability that a and b will sit next to each other?

.....

.....

.....

.....

.....

3) Two cards are drawn at random from a pack of well-shuffled cards. Find the probability that

   a) both cards are red

   b) one is a heart and another a diamond

.....

.....

.....

.....

.....

4) A bag contains 6 white and 4 red balls. One ball is drawn at random. What is the probability that it will be white?

.....

.....

.....

.....

5) 15 identical balls are distributed at random into 4 boxes numbered 1,2,3,4. Find the probability that, (a) each box contains at least 2 objects and (b) no box is empty.

.....

.....

.....

.....

6) A box contains 20 identical tickets, the tickets being numbered as 1,2,3,...., 20. If 3 tickets are chosen at random, what is the probability that the numbers on the drawn tickets will be in arithmetic progression?

.....

.....

.....

.....

17.5 THEOREMS OF PROBABILITY

In this section, we will consider the basic theorems of probability. We will denote the sub sample spaces by A, B, C ... which represent the elements of the sample space S, which are favorable to the events A, B and C.

Note that P(S) = 1 and P(A), P(B),P(C) ....lie between 0 and 1.

17.5.1 Theorem of Total Probability

If two events A and B are mutually exclusive, exhaustive and equally likely, then the occurrence of either A or B, (A U B) is given by the sum of their probability. Thus,

$P(A \cup B) = P(A) + P(B)$

This is also known as the Addition Theorem.

Proof: Let us assume that a random experiment has n possible outcomes which are mutually exclusive, exhaustive and equally likely. While m_1 of them are favorable to A, m_2 are favorable to B. By the classical definition of probability

$P(A) = m_1 / n$ and $P(B) = m_2 / n$

Since A and B are mutually exclusive and exhaustive, the number of events favorable to the event (A U B) is given by $m_1 + m_2$, therefore,

$P(A \cup B) = (m_1 + m_2) / n = (m_1 / n) + (m_2 / n) = P(A) + P(B)$ (*proved*)

17.5.1.1 Deductions from Theorem of Total Probability

1) **Theorem of Complementary Event**

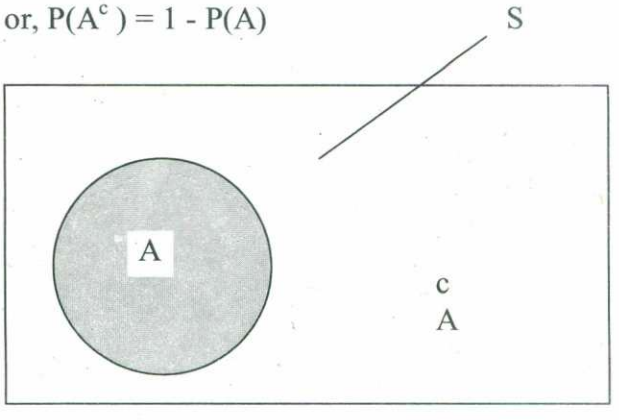
If A denotes the occurrence of the event A, then A^c (read as 'compliment of A') denotes non occurrence of the event A and

$P(A) = 1 - P(A^c)$.

Since A and A^c are mutually exclusive and exhaustive events, $S = \{A, A^c\}$. Applying the theorem of total probability we get,

$P(S) = P(A) + P(A^c) = 1$

or, $P(A^c) = 1 - P(A)$



See that this theorem is very intuitive. If the probability of getting head while tossing an unbiased coin is .5 then the probability of getting a tail is obviously .5 (1 - .5).

2) **Extension of Total Probability Theorem**

The theorem of total probability could be extended to any number of mutually exclusive events. If the events A_1 , A_2 , A_3 , ……., A_k are mutually exclusive, then the probability of occurrence of any one of them ($U^k_{i=1}$ A_i) is given by the sum of their probabilities.

$P(U^k_{i=1} A_i) = P(A_1) + P(A_2) + P(A_3) + \ldots\ldots\ldots + P(A_k)$

3) **Theorem of Total Probability with Mutually Non-exclusive Events**

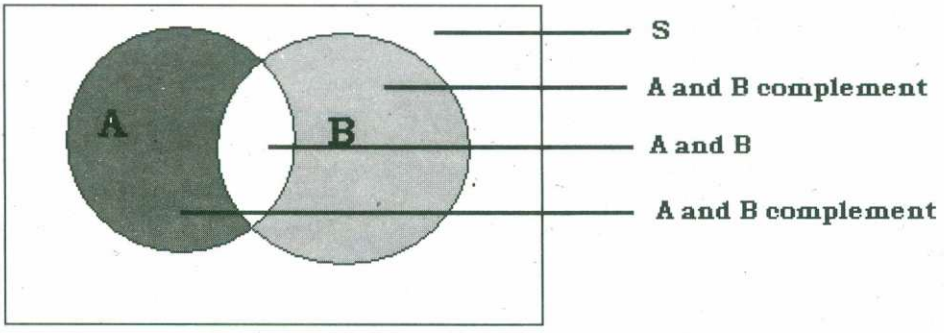
The probability of occurrence of at least one of the events A and B (which are not necessarily mutually exclusive) is given by

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The symbol '$\cap$' means 'and' i.e., $(A \cap B)$ means the occurrence of the event A and B, whereas 'U' means 'or' i.e., $(A \cup B) \Rightarrow$ the occurrence of either the event A or the event B.

Proof: The occurrence of the event $(A \cup B)$ is analogous to the occurrence of any one of the following three mutually exclusive events:

$(A \cap B^c)$, $(A^c \cap B)$ and $(A \cap B)$. In terms of Venn diagram,



Therefore, using the definition of total probability, we get

$P(A \cup B) = P(A \cap B^c) + P(A^c \cap B) + P(A \cap B)$ ..... [1]

Again, the occurrence of A is analogous to the occurrence of any one of the following two mutually exclusive events $P(A \cap B)$ and $P(A \cap B^c)$, Thus we get

$P(A) = P(A \cap B) + P(A \cap B^c)$ .....[2]

Similarly, for B

$P(B) = P(B \cap A) + P(B \cap A^c)$ .....[3]

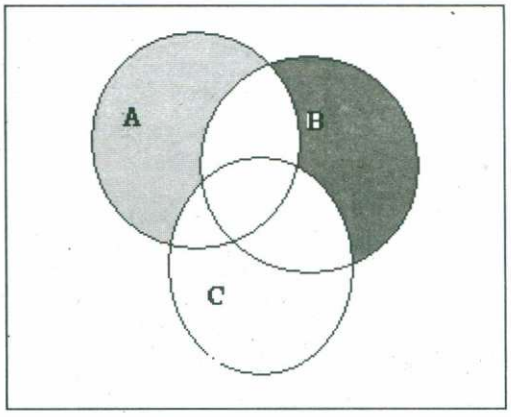
Using [1], [2], [3] we can derive that

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (*proved*).

The above result could be extended to three events A, B, C, which are not mutually exclusive

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(C \cap B) + P(A \cap B \cap C)$

In the following figure, we illustrate the situation.



In this context, we mention that there are two standard results in the theory of probability

1) $P(A \cup B) \leq P(A) + P(B)$ .....Boole's inequality

2) $P(A \cap B) \geq P(A) + P(B) - 1$ .....Bonferroni's inequality

17.5.2 Theorem of Compound Probability

The probability of occurrence of the event A and B simultaneously is given by the product of the probability of the event A and *conditional probability* of the event B given that A has actually occurred, which is denoted by P(A/B). P(A/B) is given by the ratio of the number of events favorable to the event A and B to the number of events favorable to the event A. Symbolically,

$P(A \cap B) = P(A) \times P(B/A)$.

Proof: Suppose a random experiment has n mutually exclusive, exhaustive and equally likely outcomes among which m_1, m_2 and m_{12} are favorable to the events A, B and $(A \cap B)$ respectively.

$P(A \cap B) = m_{12} / n$

$= m_1/n \times m_{12} / m_1$

$= P(A) \times P(B/A)$ (Proved).

This theorem is also known as the multiplication theorem.

17.5.2.1 Deductions from Theorem of Total Probability

The occurrence of one event, say, B may be associated with the occurrence or non-occurrence of another events say, A. This in turn implies that we can think of B to be composed of two mutually exclusive events $(A \cap B)$ and $(A^c \cap B)$. Applying the theorem of total probability

$P(B) = P(A \cap B) + P(A^c \cap B)$

$= P(A) \times P(B/A) + P(A^c) \times P(B/A^c)$... [using theorem of compound probability]

1) **Extension of Compound Probability Theorem**

The above theorem can be extended to include the cases when there are three or more events. Suppose there are three events A, B and C, then

$P(A \cap B \cap C) = P(A) \times P(B/A) \times P(C/(A \cap B))$

And so on for more than three events.

Example 2: Given $P(A) = 3/8$, $P(B) = 5/2$ and $P(A \cup B) = ¾$ find $P(A/B)$ and $P(B/A)$.

$P(A \cap B) = P(A) + P(B) - P(A \cup B) = ¼$

Therefore, $P(A/B) = P(A \cap B) / P(B) = 2/5$

and $P(B/A) = P(A \cap B) / P(A) = 2/3$

Example 3: At an examination in three courses A, B and C the following results were obtained

25% of the candidates passed in course A

20% of the candidates passed in course B

35% of the candidates passed in course C

7% of the candidates passed in course A and B

5% of the candidates passed in course A and C

2% of the candidates passed in course B and C

1% of the candidates passed in all the subject

Find the probability that a candidate got pass marks in at least one course.

$P(A) = .25$, $P(B) = .2$, $P(C) = .35$, $P(A \cap B) = .07$, $P(C \cap B) = .05$, $P(A \cap C) = .02$ and $P(A \cap B \cap C) = .01$.

Therefore, $P(A \cup B \cup C) = .25 + .2 + .35 - .07 - .05 - .02 + .01 = .67$

Check Your Progress 2

1) If $P(A) = ½$, $P(B) = 1/3$, $P(A \cap B) = ¼$ find $P(A^c)$, $P(A \cup B)$, $P(A/B)$, $P(A^c \cap B)$, $P(A^c \cap B^c)$, $P(A^c \cup B)$.

.....

.....

.....

.....

.....

2) For three events A, B, C which are not mutually exclusive, prove $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(C \cap B) + P(A \cap B \cap C)$ using Venn diagram.

.....

.....

.....

.....

3) Prove Boole's inequality and Bonferroni's inequality.

.....

.....

.....

.....

.....

17.6 CONDITIONAL PROBABILITY AND CONCEPT OF INDEPENDENCE

17.6.1 Conditional Probability

From the theorem of compound probability we can get the probability of one even, say, event B conditioned on some other event, say A. As we have discussed earlier, this is symbolically written as P(B/A). From the theory of compound probability, we know that

$P(A \cap B) = P(A) \times P(B/A)$

or, $P(B/A) = P(A \cap B) / P(A)$ provided that $P(A) \neq 0$.

Example 4: Find out the probability of getting the Ace of hearts when one card is drawn from a well-shuffled pack of cards given the fact that the card is red.

Let A denotes the event that the card is red and B denotes the event that the card is the Ace of hearts. Then clearly we are interested in finding P (B/A). From the theorem of conditional of probability

$P(B/A) = P(A \cap B) / P(A) = (1/52)/(26/52) = 1/26$

17.6.2 Concept of Independent Events

Two events A and B are said to be statistically independent if the occurrence of one event is not affected by the occurrence of another event. Similarly, several events are said to be independent, mutually independent or statistically independent if the occurrence of one event is not affected by the supplementary knowledge of the occurrence of other events. These imply that

$P(B/A) = P(B/A^c) = P(B)$

Therefore, from the theorem of compound probability, we get

$P(A \cap B) = P(A) \times P(B/A)$

$= P(A) \times P(B)$

Similarly, for three events we have the following results is that events are mutually or statistically independent

$P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$ along with

$P(A \cap B) = P(A) \times P(B)$

$P(C \cap B) = P(C) \times P(B)$

$P(C \cap A) = P(C) \times P(A)$

For more events A, B, C, D to be mutually independent following should hold:

$P(A \cap B \cap C \cap D) = P(A) \times P(B) \times P(C) \times P(D)$ along with

$P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$

$P(A \cap B \cap D) = P(A) \times P(B) \times P(D)$

$P(D \cap B \cap C) = P(D) \times P(B) \times P(C)$

$P(A \cap D \cap C) = P(A) \times P(D) \times P(C)$

$P(A \cap B) = P(A) \times P(B)$

$P(C \cap B) = P(C) \times P(B)$

$P(C \cap A) = P(C) \times P(A)$

and so on.....

Thus, two events are said to be independent if the probability of occurrence of both or all equals the product of their probabilities.

For pair wise independence of events the above should hold for any two of the events.

Deductions from the Concept of Independence

If the events A and B are independent then A^c and B^c are also independent.

Proof: Since A and B are independent

$P(A \cap B) = P(A) \times P(B)$

$P(A^c \cap B^c) = P(A \cup B)^c$ .....[De Morgan's theorem]

$= 1 - P(A \cup B)$

$= 1 - P(A) - P(B) + P(A \cap B)$

$= 1 - P(A) - P(B) + P(A) \times P(B)$

$= \{1 - P(A)\}\ \{1 - P(B)\}$

$= P(A^C) \times P(B^C)$

** *try to prove the De Morgan's theorem using Venn diagram*

Example 5: Given that $P(A) = 3/8$ and $P(B) = 5/2$ and $P(A \cap B) = ¾$, find P(B/A) and P(A/B). Are A and B independent?

Using the relationship

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$, we get

$P(A \cap B) = ¼$

Thus, the given information does not satisfy the equation

$P(A \cap B) = P(A) \times P(B)$

Therefore, A and B are not independent.

Example 6: One urn contains 2 white and 2 red balls and a second urn contains 2 white an 4 red balls

a) If one ball is selected from each urn, what is the probability that they will be of the same color?

b) If an urn is selected at random and then a ball is selected at random, what is the probability that it will be a white ball?

a) Let A denote the event that both the balls drawn from each urn are of the same color. A_1 denotes that they are white and A_2 denotes that they are red. Clearly, A_1 and A_2 are two mutually exclusive events. Applying the theorem of total probability,

$P(A) = P(A_1) + P(A_2)$

Here A_1 is a compound event formed by two independent events of drawing a white ball from each urn, Therefore, $P(A_1) = ½ \times 1/3 = 1/6$. Similarly, $P(A_2) = ½ \times 2/3 = 2/6$.

Hence, $P(A) = 1/6+2/6=½$.

b) A white can be selected in two mutually exclusive ways, when urn 1 is selected and a white ball is drawn from it (denoted by the event A) and when urn 2 is selected and a white ball is drawn from it (denoted by the event B).

P(A) = P(urn 1 is selected) × P(a white ball is drawn), because selection of urn 1 and selection of a white ball are mutually independent.

P(A) = ½×½ = ¼. Similarly,

P(B) = ½×1/3 = 1/6

Using the theorem of total probability,

P(drawing a white ball from an urn) = P(A) + P(B) = 5/12.

Check Your Progress 3

1) A salesman has 50% chance of making a sale. If two customers enter the shop, what is the probability that the salesman will make a sale?

.....

.....

.....

.....

.....

2) If the events A and B are independent, then show that A^c, B^c, A and B are pair wise independent.

.....

.....

.....

.....

.....

3) If A and B are mutually exclusive events, then show that P(A/AUB) = P(A)/P(A) + P(B).

.....

.....

.....

.....

.....

4) What is the difference between mutually independent random variables and pair wise independent random variables?

.....

.....

.....

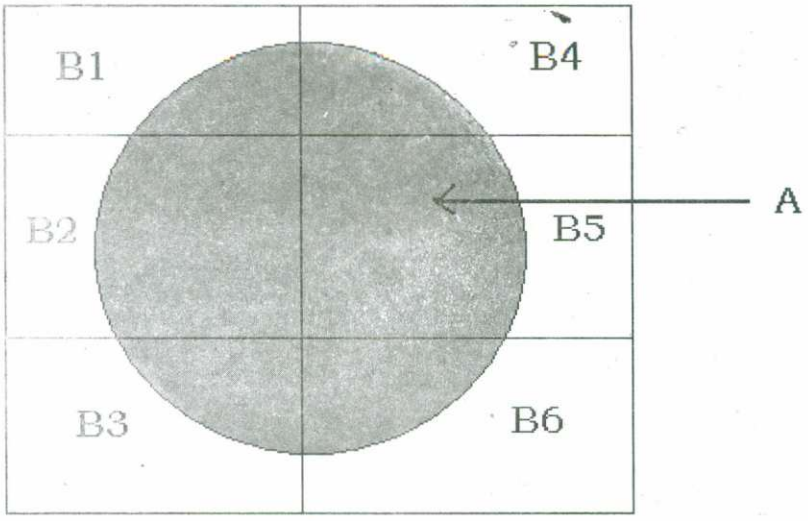
.....

.....

.....

17.7 BAYES' THEOREM AND ITS APPLICATION

Suppose an event A can occur if and only if one of the mutually exclusive events B_1, B_2, B_3,....., B_n occurs. If the unconditional probabilities $P(B_1)$, $P(B_2)$, $P(B_3)$,....., $P(B_n)$ are known and the conditional probabilities are $P(A/B_1)$, $P(A/B_2)$,



$P(A/B_3)$,....., $P(A/B_n)$ are also known. Then the conditional probability $P(B_i/A)$ could be calculated when A has actually occurred.

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)\, P(A/B_i)$$

$$P(B_i/A) = P(B_i \cap A) / P(A) = P(A/B_i) \times P(A) / \sum_{i=1}^{n} P(B_i)\, P(A/B_i), \text{ therefore}$$

$$P(B_i/A) = P(A/B_i) \times P(A) / \sum_{i=1}^{n} P(B_i)\, P(A/B_i)$$

This is known as Bayes' theorem. This is a very strong result in the theory of probability. An example will illustrate the theorem more vividly.

Example: Two boxes contain respectively 2 red and 2 black balls and 2 red and 4 black balls. One ball is transferred from one box to another and then one ball is selected from the second box. If it turns out to be black, what is the probability that the transferred ball was red?

B_1: the transferred ball was red.

B_2 : the transferred ball was black.

A: the ball selected from the second box is black.

$P(B_1) = ½$

$P(B_2) = ½$

$P(A/B_1) = 3/7$

$P(A/B_2) = 5/7$

$P(B_1/A) = P(B_1) \times P(A/B_1) / P(B_1) \times P(A/B_1) + P(B_2) \times P(A/B_2)$

$= ½ \times 3/7 / (½ \times 3/7 + ½ \times 5/7) = 3/8$

MEC-103-5-15 F-7K

Check Your Progress 4

1) In a bulb factory there are three machines a, b and c. They produce 25%, 35% and 40% of total product. Of their output 5, 4 and 2 per cent respectively are defective. If one bulb is selected at random, what is the probability that it was produced by the machine c?

.....

.....

.....

.....

.....

2) We have two coins. The first coin is fair with probability of head= ½, but the second coin is loaded with head. Therefore, probability of getting head in the second coin = 2/3. Suppose we pick one coin at random, we toss it, and the head shows. Find the probability that we picked the fair coin.

.....

.....

.....

.....

.....

17.8 MATHEMATICAL EXPECTATIONS

Suppose a random experiment has n mutually exclusive and exhaustive outcomes. Let the variable x can take values $x_1, x_2, x_3, \ldots\ldots, x_n$ with probabilities $p_1, p_2, p_3, \ldots\ldots, p_n$. Then the mathematical expectation or the expected value of the variable is defined as the weighted sum of the values of the variable the weights being the probabilities of the values of the variable. Expected value of a variable 'x' is denoted by E(x). Therefore,

$$E(x) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots\ldots + x_np_n = \sum_{i=1}^{n} X_i\, p_i$$

provided that $\sum_{i=1}^{n} p_i = 1$

If E(x) = m, the mathematical expectation of the variable $(x - m)^2$ is known as the variance of the variable x, which is denoted by Var(x).

$$Var(x) = E(x - m)^2 = (x_1 - m)^2 p_1 + (x_2 - m)^2 p_2 + (x_3 - m)^2 p_3 + \ldots\ldots + (x_n - m)^2 p_n$$

$$= \sum_{i=1}^{n} (x_i - m)^2 p_i$$

We can show that $Var(x) = E(x - m)^2 = E(x^2) - [E(x)]^2$. Remember, in our notation, E(x) = m. The square root of Var(x) is called the standard deviation of the variable x.

If g(x) is any function of the variable x, defined for every value of the variable x, then the expected value of the function g(x), denoted by E[g(x)] is given by

$$E[g(x)] = g(x_1)\, p_1 + g(x_2)p_2 + g(x_3)p_3 + \ldots\ldots + g(x_n)p_n = \sum_{i=1}^{n} X_i - m)^2\, p_i$$

Mathematical expectation of a variable is analogous to weighted Arithmetic Mean of the variable, say, x. In a frequency distribution, the relative frequencies (Class frequency / Total frequency) of a variable could be thought as the probability that the variable will take that value, i.e. $p_i = f_i / N$,

where, p_i: Probability that the variable x will take the value x_i

f_i: frequency of occurrence of x_i

N: Total frequency $\sum_{i=1}^{n} f_i$

As we know weighted arithmetic mean $= \frac{\sum_{i=1}^{n} X_i f_i}{\sum_{i=1}^{n} f_i}$

$= \frac{\sum_{i=1}^{n} X_i f_i}{N}$

$= \sum_{i=1}^{n} X_i p_i$

= E(x)

= Expected value of the variable x.

Example: What is the mathematical expectation of the number of points when an unbiased die is thrown?

Let the variable denote the number of points when a die is thrown. Therefore, x can take values 1, 2, 3, 4, 5 and 6. The probability of realization of all these values is same, viz.,1/6, therefore,

E(x) = 1/6 (1+2+3+4+5+6) = 3.5

Theorem 1

The mathematical expectation of sum of several random variables is equal to the sum of the mathematical expectation of the random variables.

E(a+b+c+d+.....) = E(a)+E(b)+E(c)+E(d)+....., where a, b, c, d represent random variables.

We will prove the theorem for two random variables and the result could be extended to any number of random variables.

Let, x and y are two random variables; x can take the values $x_1, x_2, x_3, \ldots, x_n$ and y can take the values $y_1, y_2, y_3, \ldots, y_m$, and p_{ij} is the probability of the event that $x=x_i$ and $y=y_j$. Now (x + y) is a new random variable and it takes the value $(x_i + y_j)$ with probability p_{ij} or, $P(x=x_i \text{ and } y=y_j) = p_{ij}$. Using the definition of mathematical expectation

$E(x + y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i + y_i)p_{ij}$ [we take double summation as i can take n values and j can take m values]

$E(x + y) = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i p_{ij} + y_j p_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_{ij}$

$= \sum_{i=1}^{n} x_i \sum_{j=1}^{m} p_{ij} + \sum_{j=1}^{m} y_j \sum_{i=1}^{n} p_{ij}$ [we could write this as x_i is constant with respect to variations in j]

The following diagram will explain how we could write the above. In this context, we define marginal probability of a variable given the other variable takes some specific value. Suppose n= 9 and m= 8, i.e., the variables x and y take 9 and 8 values respectively. From the diagram, it is easy to see the distribution of the variables and their probabilities, where the symbols have their usual meanings.

Marginal probability of y takes the value $y_j = \sum_{i=1}^{9} p_{ij} = p_{0j}$

y → / x ↓	y1	y2	y3	y4	y5	y6	y7	y8	Marginal Probability of x ↓
x1	p11	p12	p13	p14	p15	p16	p17	p18	p10
x2	p21	p22	p23	p24	p25	p26	p27	p28	p20
x3	p31	p32	p33	p34	p35	p36	p37	p38	p30
x4	p41	p42	p43	p44	p45	p46	p47	p48	p40
x5	p51	p52	p53	p54	p55	p56	p57	p58	p50
x6	p61	p62	p63	p64	p65	p66	p67	p68	p60
x7	p71	p72	p73	p74	p75	p76	p77	p78	p70
x8	p81	p82	p83	p84	p85	p86	p87	p88	p80
x9	p91	p92	p93	p94	p95	p96	p97	p98	p90
Marginal Probability of y →	p01	p02	p03	p04	p05	p06	p07	p08	1

Similarly, marginal probability of x takes the value x_i

$= \sum_{j=1}^{8} p_{ij} = p_{i0}$

In the above situation, $E(x) = \sum_{i=1}^{9} x_i p_{0i}$ and $E(y) = \sum_{i=1}^{n} y_i p_{j0}$

Additionally, we can also derive the conditional distribution of the variables from the above diagram. One example will elucidate it. The conditional distribution of x given $y = y_1$ is as follows:

	y = y1	Conditional probability distribution of x given y = y1
x1	p11	p11 / p01
x2	p21	p21 / p01
x3	p31	p31 / p01
x4	p41	p41 / p01
x5	p51	p51 / p01
x6	p61	p61 / p01
x7	p71	p71 / p01
x8	p81	p81 / p01
x9	p91	p91 / p01
Total	p01	1

Therefore, $E(x + y) = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} p_{ij} + \sum_{j=1}^{m} y_j \sum_{i=1}^{n} p_{ij}$

$= \sum_{i=1}^{n} x_i p_{i0} + \sum_{j=1}^{m} y_j p_{0j} = E(x) + E(y)$ (proved).

Theorem 2

The mathematical expectation of product of several independent random variables is equal to the product of the mathematical expectation of the random variables.

We retain the symbols of the previous theorem and additionally, we assume that the variables x and y are independent, i.e., occurrence of any one of the event has no impact on the occurrence of the other. Let the variable x take the value x_i with probability p_i and the variable y takes the value y_j with probability q_j . Since x and y are independent

$P(x=x_i \text{ and } y=y_j) = p_i \times q_j$.

The theorem states that

$E(x.y) = E(x) \times E(y)$.

Proof: Using the definition of mathematical expectation

$$E(x.y) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p(x=x_i \text{ and } y=y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j (p \times q_j)$$

Summing over j and keeping i constant,

$= \sum_{i=1}^{n} x_i p_i \times \sum_{j=1}^{m} y_j q_j) = E(x) \times E(y)$.....(proved).

This theorem could be extended to any number of variables.

Corollary 1: If m is the expected value of the variable x, then

$E(x - m)^2 = E(x^2) - [E(x)]^2$.

$E(x - m)^2 = \sum_{i=1}^{n}(x_i - m)^2 p_i = \sum_{i=1}^{n}(x_i^2 - 2.x^i.m + m^2) p_i = \sum_{i=1}^{n} x_i^2 \times p_i - 2.m. \sum_{i=1}^{n} x_i \times p_i + m^2$

$\sum_{i=1}^{n} p_i = E(x^2) - 2.m.E(x) + m^2 = E(x^2) - 2.m.m + n.m^2 = E(x^2) - 2.m.E(x) + m^2 =$

$E(x^2) - 2.m.m + m^2 = E(x^2) - m^2 = E(x^2) - E(x)^2$ (proved).

$E(x - E(x))^2$ is called the variance of the variable x and it is denoted by Var (x).

Corollary 2: If there are n random variables $x_1, x_2, x_3, \ldots\ldots, x_n$ each having mean $m_1, m_2, m_3, \ldots\ldots, m_n$,

$$\text{Var}(x_1 + x_2 + x_3 + \ldots\ldots + x_n) = \sum_{i=1}^{n} \text{Var}(x_i) + 2 \sum_{i=1}^{n} \text{Cov}(x_i, x_j)$$

$\text{Cov}(x_i, x_j) = E(x_i - m_i)(x_j - m_j)$ when $\neq j$.

$\text{Var}(x_1 + x_2 + x_3 + \ldots\ldots + x_n) = E[(x_1 + x_2 + x_3 + \ldots\ldots + x_n) - (m_1 + m_2 + m_3 + \ldots\ldots + m_n)]^2$

$= \sum_{i=1}^{n} E(x_i - m_i) + 2 \sum_{i=1}^{n} E(x_i - m_i)(x_j - m_j)$

The first term in the above equation is Variance of the variable x_i whereas the second term is the $\text{Cov}(x_i, x_j)$. Therefore,

$$\text{Var}(x_1 + x_2 + x_3 + \ldots\ldots + x_n) = \sum_{i=1}^{n} \text{Var}(x_i) + 2\sum_{i \neq j}^{n} \text{Cov}(x_i, x_j)$$

If the variables are mutually independent then the covariance term in the above expression is zero (since mutual independence rules out joint occurrence).

Check Your Progress 5

1) A man purchases a lottery ticket. He may win first prize of Rs.10,000 with probability .0001 or the second prize of Rs. 4,000 with probability .0004. On an average, how much he can expect from the lottery?

.....

.....

.....

.....

.....

2) A box contains 4 white balls and 6 black balls. If 3 balls are drawn at random, find the mathematical expectation of the number of white balls.

.....

.....

.....

.....

.....

3) If $y = a + b.x$, where a nd b are constants, prove that $E(y) = a + bE(x)$.

.....

.....

.....

.....

.....

17.9 LET US SUM UP

The term probability in a very crude sense implies the chance factor and it is used frequently where there is uncertainty about something. Mathematicians from the very early ages tried endlessly to build up a structure under which this uncertain thing called probability could be analysed. In this unit, we have discussed the various methods frequently used to determine the probability of various non-deterministic experiments. Using these techniques, we can determine the probabilities of many day-to-day life uncertain events. Mathematical expectation introduced at the end of this unit is nothing but the mean of a random variable. It is quite useful in understanding the nature of a random variable

17.10 KEY WORDS

Sample Space: The collection of all possible outcomes of an experiment is called the sample space.

Events: When some of the elements of the sample space of a random experiment satisfy a particular criterion, we call it an event.

Mutually Exclusive Events: Events are said to be mutually exclusive when no two or more of them can occur simultaneously.

Mutually Exhaustive Events: Events are exhaustive if at least one of them necessarily occurs.

Independent Events: Events are said to be independent of each other if occurrence of one event is not affected by the occurrence of the others. If events A and B are independent then $P(A \cap B) = P(A) \times P(B)$.

Conditional Events: If two events A and B are mutually exclusive and if it is known that event b has already taken place; the probability of A is known as the conditional probability of A given B. Symbolically $P(A/B) = P(A \cap B) / P(B)$.

Bayes' Theorem: If an event A occurs in conjunction with one of the n mutually exclusive and collectively exhaustive events $E_1, E_2, E_3 \ldots E_n$ and if A actually occurs, then the probability that it was preceded by a particular event E_i (I = 1, 2, 3……n) is given by

$$P(B_i/A) = P(A/B_i) \times P(A) / \sum_{i=1}^{n} P(B_i)\, P(A/B_i)$$

Marginal Probability: In a bivariate distribution the probability that X will assume a given value X whatever the value of Y is called the marginal probability of Y. Similarly, one can define the marginal distribution of Y.

Mathematical Expectation: If the variable x can take values $x_1, x_2, x_3, \ldots\ldots, x_n$ with probabilities $p_1, p_2, p_3, \ldots\ldots, p_n$, then the mathematical expectation or the expected value of the variable is defined as the weighted sum of the values of the variable the weights being the probabilities of the values of the variable. Expected value of a variable 'x' is denoted by E(x).

$$E(x) = x_1p_1 + x_2p_2 + x_3p_3 + \ldots\ldots + x_np_n = \sum_{i=1}^{n} x_ip_i,$$

provided that $\sum_{i=1}^{n} p_i = 1$.

17.11 SOME USEFUL BOOKS

Das.N.G. (1996), *Statistical Methods*, M.Das & Co.(Calcutta)

Feller, W. (1968), *An Introduction to Probability and Its Applications*, Vols. 1&2, 3rd ed. New York: Wiley.

Freund J.E. (2001), *Mathematical Statistics*, Prentice Hall of India.

Goldberg, S. (1986), *Probability: An Introduction*, New York: Dover.

Hoel, P (1962), *Introduction to Mathematical Statistics*, Wiley John & Sons, New York.

Hoel, Paul G. (1971), *Introduction to Probability Theory*, Universal Book Stall, New Delhi.

17.12 ANSWER OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) If the three children were born on three different days, the first child may born in any one of 365 days, the second children has to born on anyone of the remaining 364 days and the third children has to born on anyone of the remaining 363 days. Therefore, the probability that the three children will be born on 3 different days in a year is 365.364.363/365.365.365.

2) Five persons can sit in a row in 5! = 5.4.3.2.1 = 120 ways. Considering a and b together they can arrange among themselves in four! 2!= 48 ways. Therefore, the required probability is = 48/120= .4.

3) Two cards can be drawn from a pack of 52 cards in $^{52}C_2$= 1326 ways. These outcomes are mutually exclusive exhaustive and equally likely.

a) The number of cases favorable to both the cards are red is $^{26}C_2$= 325. Therefore, that probability of drawing both red cards = 325/1326.

b) One heart and one diamond can be drawn in13.13= 169 ways. Therefore, probability of drawing one heart and one diamond = 169/1326.

4) One white ball could be drawn in $^{6}C_1$ = 6 ways and one ball can be drawn in $^{10}C_1$ = 10 ways. Therefore, the probability of drawing one white ball = 6/10 = .6.

5) The total number of ways of distributing n identical objects into r compartments is given by the formula $^{n+r-1}C_{r-1}$ Using the formula we can find out that there are 816 mutually exclusive exhaustive and equally likely ways of distributing 15 identical objects into 4 numbered boxes.

   a) If each box is to contain at least 2 objects, we place 2 objects in each box and then the remaining 7 objects could be distributed among the 4 boxes. Using the above formula we get there are 120 ways of doing that. Therefore, the required probability is 5/34.

   b) If the number box has to be empty then it means there should be at least one object in each box. We first distribute 1 object in each box. Then the remaining 11 objects could be distributed in 364 ways (using the above mentioned formula). Therefore, the required probability is given by 91/204.

6) Total number of possible equally likely exhaustive and exclusive ways of choosing 3 tickets is given by $^{20}C_3$ = 1140. The three numbers will be in A.P if the difference between the numbers is either 1 or 2 or 3 .....9. If the difference is 1 then 18 sets of numbers are possible, (123, 234, 345, ...181920). Similarly, if the difference is 2 then 16 sets of numbers are possible, (135, 246, 357, ...161820). Thus, we can find the total number of sets of 3 numbers all of them being in A.P is 90. and the required probability is 3/38.

Check Your Progress 2

1) $P(A^c) = ½$,$P(AUB) = 7/12$, $P(A/B) = 3/4$, $P(A^c \cap B) = 1/12$, $P(A^c \cap B^c) = 5/12$, $P(A^c \cup B) = 3/4$.

2) Do yourself using Sub-section 17.6.1.3.

3) Do yourself using Sub-section 17.6.1.3.

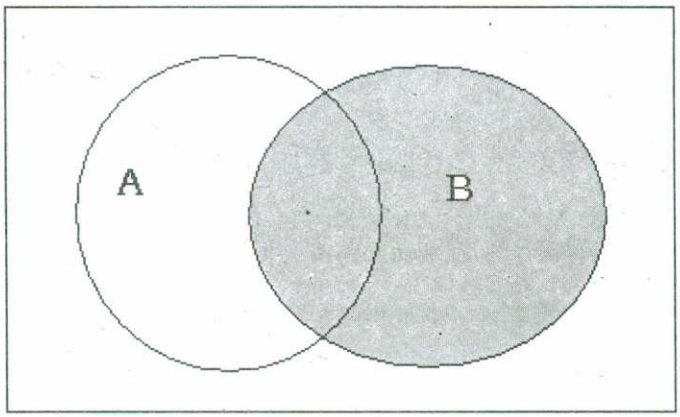
Check Your Progress 3

1) A: The salesperson makes sale to the first customer.

B: The salesperson makes sale to the second customer.

$P(A) = P(B) = ½$

(A U B): the salesperson will make a sell.

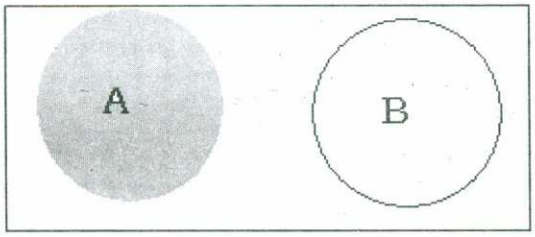


$P(A \cup B) = 1 - P(A^c \cap B^c) = 1 - P(A^c)(B^c)$ [as events A and B are independent]

$= 1 - ½ . ½ = ¾$

2) Do yourself using Sub-section 17.7.2.

3)



$P(A/AUB) = P(A \cap (A \cup B))/P(A) + P(B)$. [$P(A \cup B) = P(A) + P(B)$ theorem of total probability]

$= P((A \cap A) \cup (A \cap B)) = P(A \cup (A \cap B)) = P(A) + P(A \cap B) - P(A \cap (A \cap B)) = P(A) + 0 - 0 = P(A)$

Therefore $P(A/AUB) = P(A)/P(A) + P(B)$.

4) Do yourself using Sub-section 17.7.2

Check Your Progress 4

1) A: Selected bulb is defective.

B_1: Selected bulb has been produced by machine 'a'

B_2: Selected bulb has been produced by machine 'b'

B_3: Selected bulb has been produced by machine 'c'

Bi	P(Bi)	P(A/Bi)	P(Bi). P(A/Bi)
B1	0.25	0.05	0.0125
B2	0.35	0.04	0.014
B3	0.4	0.02	0.008
Total	1		0.0345

From Bayes' theorem

$P(B_3/A) = P(A/B_3) \times P(A) / \sum_{i=1}^{n} P(B_i)\, P(A/B_i) = .008 / .0345 = 16/69.$

2) Follow the same method. Answer = 3/7.

Check Your Progress 5

1) The calculations for mathematical expectation are provided below

X	P(x)	x. P(x)
0	0.9995	0
4000	0.0004	1.6
10000	0.0001	1
Total		2.6

E(x) = 2.6 Rs.

2) Probability of drawing '0' white ball = ${}^4C_0\,{}^6C_3/{}^{10}C_3 = 1/6$

Probability of drawing '1' white ball = ${}^4C_1\,{}^6C_2/{}^{10}C_3 = ½$

Probability of drawing '2' white ball = ${}^4C_2\,{}^6C_1/{}^{10}C_3 = 3/10$

Probability of drawing '3' white ball = ${}^4C_3\,{}^6C_0/{}^{10}C_3 = 1/30$

Fill up the following blank table to get the required expectation

X	p(x)	x.P(x)
0		
1		
2		
3		
Total		

3) Do yourself using Section 17.8.

17.13 EXERCISES

1) Give the classical definition of probability, what do you think could be its limitations?

2) State truth value (whether true or false) of each of the following statements

i) $P(A \cup B) + P(A \cap B) = P(A) + P(B)$

ii) $P(A \cap B) = P(A \cup B) + P(\bar{A} \cup B) + P(A \cup \bar{B})$

iii) $P(A/B) \times P(B/A) = 1$

iv) $P(A/B) \leq P(A)/P(B)$

v) $P(A/B) = 1 - P(A/B)$

3) The nine digits 1, 2, 3..., 9 are arranged in random order to form a nine digit number. Find the probability that the numbers 2, 4, 5 appear as neighbor in the order they are mentioned?

4) Four dice are thrown. Find the probability that the sum of the numbers appearing in the four dice is 20.

5) There are three persons in a group. Find the probability that

   a) all of them have different birthdays;

   b) at least two of them have the same birthday;

   c) exactly 2 of them have the same birthday.

6) An urn contains 7 red and 5 white balls. 4 balls are drawn at random. What is the probability that all of them are red and 2 of them are red and 2 are white?

7) The incidence of a certain epidemic is such that on an average 20% of the people are suffering from it. If 10 people are selected at random find the probability that exactly 2 of them suffer from the disease?

8) If a person gains or looses an amount equal to the number appearing when an unbiased die is thrown once according to the number is even or odd, how much money he can expect in the long run from the game?

9) Ram and Rahim play for a prize of Rs. 99. The prize is to be won by the player who first throws a 3 with a single die. Ram throws first and if he fails Rahim throws it and if Rahim fails Ram throws it again and this process continues. Find their respective expectations.

10) The probability that an assignment will be finished in time is 17/20. The probability that there will be a strike is ¾. The probability that an assignment will be finished in time if there is no strike is 14/17. Find the probability that there will be strike or the job will be finished in time.

11) If $P(A \cap B \cap C) = 0$, show that $P[(A \cup B)/ C] = P(A/C) + P(B /C)$.

NOTES

MPDD/IGNOU/P.O.7K/November 2018 (Reprint)

ISBN : 978-93-86375-16-2