PG EC S4-01 (DSC)

TWO YEAR MASTER'S DEGREE COURSE

(CBCS in Accordance with NEP 2020)

SEMESTER IV

STATISTICAL METHODS

ECONOMICS

Surya Kumar Bhuyan School of Social Sciences

KRISHNA KANTA HANDIQUI STATE OPEN UNIVERSITY

Subject Experts

Professor Madhurjya Prasad Bezbaruah, Dept. of Economics, Gauhati University
Professor Rajshree Bedmatta, IIT Guwahati
Professor Ratul Mahanta, Gauhati University
Dr. Gautam Mazumdar, Dept. of Economics, Cotton University, Guwahati

Course Co-ordinator: Dr. Parag Datta, KKHSOU

SLM Preparation Team

UNITS	CONTRIBUTORS
1, 2, 4 & 5	**Dr. Pranjal Sarma**, Department of Statistics L. C. B. College, Guwahati
3, 11	**Dr. Pranjal Sarma**, Department of Statistics L. C. B. College, Guwahati, & **Dr. Bhaskar Sarmah,** KKHSOU
6, 7, 8, 9 & 10	**Dr. Rijusmita Sarma**, HoD, Department of Statistics L. C. B. College, Guwahati
12	**Bhaskar Sarmah**, KKHSOU **Dr. Rupam Barman,** Tezpur University
13, 14	**Dr. Mitali Deka,** NERIM, Guwahati

Editorial Team

Content	:	**Dr. Parag Dutta,** Associate Professor in Economics, KKHSOU (Units 1 - 12) **Professor Tarakeswar Choudhary** (Retired), Cotton College, Guwahati (Units 13-14)
Structure, Format	:	**Dr. Bhaskar Sarmah,** Assistant Professor in Economics, KKHSOU
& Graphics	:	**Dr. Harekrishna Deka,** Assistant Professor in Mathematics, KKHSOU

March, 2025

ISBN: 978-93-89955-67-5

This Self Learning Material (SLM) of the Krishna Kanta Handiqui State Open University is made available under a Creative Commons Attribution-Non Commercial-Share Alike 4.0 License (international): http://creativecommons.org/licenses/by-nc-sa/4.0/

Printed and published by Registrar on behalf of the Krishna Kanta Handiqui State Open University

Headquarter : Patgaon, Rani Gate, Guwahati - 781017

City Office : Resham Nagar, Khanapara, Guwahati-781022; Website: www.kkhsou.in

The University acknowledges with thanks the financial support provided by the **Distance Education Bureau**, **UGC** for the preparation of this study material.

CONTENTS

		Pages
UNIT 1:	**Measures of Central Tendency** Arithmatic Mean; Median; Mode; Geometric Mean; Harmonic Mean; Weighted Average	**7–32**
UNIT 2:	**Measures of Dispersion** Range; Quartile Deviation; Mean Deviation; Standard Deviation, Standard Deviation; Coefficient of Variation	**33–54**
UNIT 3:	**Measurement of Economic Inequality** Pareto's Law of Income Distribution; Log normal Distribution (Concept only); Lorenz Curve; Coefficient of Inequality : Gini Coefficient	**55–65**
UNIT 4:	**Correlation** Correlation Analysis; Types of Correlation; Methods of Measuring Correlation; Rank Correlation Coefficient; Tied or Repeated Ranks	**66–88**
UNIT 5:	**Regression Analysis** Regression; Two-Variable Regression Model (Least Square Method); Properties of Regression, Coeffieicents; Correlation vs. Regression	**89–111**
UNIT 6:	**Probability : Basic Ideas** Definition of the Basic Terms; Classical or Mathematical or Apriori Definition of Probability; Addition Theorems or Theorems on Total Probability; The Theorem of Compound Probability or the Multiplicative Law of Probability; Baye's Theorem (Concept only)	**112–129**
UNIT 7:	**Expectations and Moments** Random Variable: Definition or Random Variable, Discrete Random Variable, Continuous Random Variable; Probability Distribution: Probability Mass Function, Probability Density Function; Mathematical Expectation and Variance: Mathematical Expectation, Properties of Mathematical Expectation, Variance of Random Variable, Properties of Variances; Moments; Skewness: Karl Pearson's Co-efficient of Skewness, Boweley's Co-efficient of Skewness; Kurtosis: Measures of Kurtosis	**130–162**

UNIT 8: Theoretical Probability Distributions (Discrete Variable) **163–192**

Moment Generating Function: Definition, Generation of Moments, Properties of Moment Generating Function; Binomial Distribution: Derivation of Binomial Distribution, Definition, Moments of Binomial Distribution, Moment Generating Function of Binomial Distribution, Fitting of Binomial Distribution, Properties of Binomial Distribution; Poisson Process: Poisson Distribution as a Limiting Case of Binomial Distribution, Moments of Poisson Distribution, Moment Generating Function of Poisson Distribution, Fitting of Poisson Distribution, Properties of Poisson Distribution, Application of Poisson Distribution

UNIT 9: Theoretical Probability Distributions (Continuous Variable) **193–211**

Normal Distribution: Definition, Standard Normal Variate, Properties of Normal Distribution, Area under Normal Probability Curve, Importance of Normal Distribution; Weak Law of Large Numbers: Statement, Application of Weak Law of Large Numbers; Central Limit Theorem: Statement, Application of Central Limit Theorem

UNIT 10: Sampling and Population **212–238**

Concept of Population and Sample: Basic Principles of Sample Survey, Advantages of Sampling over Complete Enumeration, Different Steps in a Sample Survey, Errors in Surveys (Non-Sampling and Sampling Errors), Types of Sampling, Selection of Random Samples; Simple Random Sampling: Estimate of Population; Stratified Random Sampling: Advantage and Disadvantages of Stratified Random Sampling, Allocation in Stratified Random Sampling; Systematic Random Sampling: Linear Systimatic Sampling, Circular Systematic Sampling, Advantages and Disadvantages of Systematic Sampling, Uses of Systematic Sampling; Cluster Sampling: Advantages and Disadvanges of Cluster Sampling, Optimum Cluster size

UNIT 11: Characterisitcs of a Good Estimator **239–251**

Estimators and Estimates; Types of Estimation; Criteria for a Good Estimator: Unbiasedness, Consistency, Efficiency

UNIT 12: Index Numbers **252–274**

Index Numbers: Concept and Uses; Construction of Simple Index Numbers; Construction of Weighted Index Numbers: Laspeyres' Index Number,

Paasche's Index Number, Fisher's Index Number; Test of Adequacy of Index Numbers: Time Reversal Test, Factor Reversal Test, Circular Test; Base Shifting, Splicing and Deflating; Index of Industrial Production (Concept only)

UNIT 13: Testing of Hypothesis **275–304**

Statistical Hypothesis: Null Hypothesis and Alternative Hypothesis; Errors in Hypothesis Testing, Level of Significance and Critical Region; One - Tailed and Two-Tailed Tests; General Procedure for Hypothesis Testing; Testing of Hypothesis in case of Large Samples: Hypothesis Testing for Single Population Mean, Hypothesis Testing for Single Population Proportion, Hypothesis Testing for Difference between Two Population Means, Hypothesis Testing for Difference between Two Population Means

UNIT 14: Test of Significance **306–350**

The Chi-Square (χ^2) Distribution; Applications of χ^2 Test: χ^2 Test of Goodness of Fit, χ^2 Test for Independence of Attributes, χ^2 Test for a Specified Population Variance, χ^2 Test of Homogeneity; Testing of Hypothesis in case of Small Samples (t-Test): Characteristics of t-distribution, Applications of t-distribution, Hypothesis Testing for Single Population Mean, Hypothesis Testing for Difference between Two Population Means, Paired t-test for Difference between Two Population Means

References (For All Units of the Course) **251**

COURSE INTRODUCTION

This course entitled *'Statistical Methods'* is common to both MA in Economics programme offered by this University. This is the third course of the 4th Semester curriculum. As statistical tools are widely used in various economic theories and illustrations; this course has been designed to offer us the basic concept of the subject Statistics and their economic applications. This course comprises 14 units.

The course begins with an introduction to measures of central tendency in Unit 1, followed by an exploration of measures of dispersion in Unit 2. Economic inequality is the focus of Unit 3, where concepts such as Pareto's law of income distribution, the Lorenz curve, and the Gini coefficient are discussed. Correlation is examined in Unit 4, covering scatter diagrams, types of correlation, the correlation coefficient, and different measurement methods. In Unit 5, the discussion shifts to regression, including topics like linear regression, lines of regression, and regression equations. Probability is introduced in Unit 6, with expectation and moments explored in Unit 7. Theoretical probability distributions for discrete variables are covered in Unit 8, while Unit 9 extends the discussion to continuous variables. Sampling and population are discussed in Unit 10, with various sampling techniques explained. Unit 11 focuses on the characteristics of a good estimator, highlighting key properties that define an effective estimator. The construction and common applications of index numbers are explained in Unit 12. Hypothesis testing is covered in Unit 13, where the null and alternative hypotheses, along with the process for large samples, are discussed. Concluding the course, Unit 14 explores tests of significance, emphasizing hypothesis testing for small samples ($n<30$) and chi-square tests.

While going through a unit, you will notice some along-sode boxes, which have been included to help you know some of the difficult, unseen terms. Some 'ACTIVITY' (s) has been included to help you apply your own thoughts. Again, we have included some relevant concepts in 'LET US KNOW' along with the text. At the end of sections, you will get 'CHECK YOUR PROGRESS' questions. These have been designed to self-check your progress of study. It will be better if you solve the problems put in these boxes immediately after you go through the sections of the units and then match your answers with 'ANSWERS TO CHECK YOUR PROGRESS' given at the end of each unit.

UNIT 1: MEASURES OF CENTRAL TENDENCY

UNIT STRUCTURE

1.1 Learning Objectives
1.2 Introduction
1.3 Arithmatic Mean
1.4 Median
1.5 Mode
1.6 Geometric Mean (GM)
1.7 Harmonic Mean (HM)
1.8 Weighted Average
1.9 Exercises
1.10 Let Us Sum Up
1.11 Further Reading
1.12 Answers to Check Your Progress
1.13 Model Questions

1.1 LEARNING OBJECTIVES

After going through this unit, you will able to:

- give the different measures of central tendency
- write the computational tichniques of different measures of central tendency
- state the properties and respective limitations of these measures.

1.2 INTRODUCTION

Summarisation of the data is a necessary function for any statistical analysis and interpretation. Hence, as a first step, the huge mass of unorganised data are summarised in the form of tables and frequency distributions. Further, to bring out different characteristics of the data into sharp focus, these tables and frequency distributions are to be summarised again using different statistical tools. A measure of central tendency or an avrage is a statistical tool that is used for summarisation of data for statistical analysis.

Definition of Central Tendency

In many a case, like the distributions of height, weight, marks, profit, wage and so on, it has been noted that starting with rather low frequency, the class frequency gradually increases till it reaches its maximum somewhere near the central part of the distribution and after which the class frequency steadily falls to its minimum value towards the end. Thus, central tendency may be difined as the tendency of a given set of observations to cluster around a single central or middle value and the single value that represents the given set of observations is described as a measure of central tendency or average. Hence, it is possible to condense a vast mass of data by a single representative value. Furthermore, the central tendency also facilitates us in providing a basis for comparison between different distributions. Following are the different measures of central tendency :

(i) Arithmetic Mean (AM)

(ii) Median (Me)

(iii) Mode (Mo)

(iv) Geometric Mean (GM)

(v) Harmonic Mean (HM)

Criteria for an Ideal Measure of Central Tendency

Following are the criteria for an ideal measure of central tendency:

(i) It should be properly and unambignously defined.

(ii) It should be easy to compute.

(iii) It should be easy to understand.

(iv) It should be based on all the observations.

(v) It should be capable of further mathematical treatment.

(vi) It should be least affected by the presence of extreme observations.

(vii) It should not be much affected by the fluctuation of sampling.

1.3 ARITHMATIC MEAN

For a given set of observations, the Aritnmatic Mean (AM) may be defined as the sum of all the observations to be divided by the number of observations. Thus, if a variable X assumes n values $x_1, x_2, ..., x_n$, then the AM of X, to be denoted by $\overline{X}$, is given by,

$$\bar{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i \text{ or } \frac{1}{n}\Sigma x$$

In case of a simple frequency distribution relating to an attribute, we have

$$\bar{X} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n}$$

$$= \frac{1}{\sum_{i=1} fi}\sum_{i=1}^{n} f_i x_i$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i x_i \text{ or } \frac{1}{N}\Sigma fx$$

Assuming the observation x_i occurs f_i times, i = 1, 2, . . ., n and $N = \sum_{i=1}^{n} f_i = \Sigma f$ = total frequency.

In case of grouped frequency distribution also we may use the above formula with x_i as the mid value of the i^{th} class interval, on the assumption that all the values belonging to the i^{th} class interval are equal to x_i.

Assumed Mean Method :

This method is used when the magnitude of individual observations is large. The method is helpful in the simplification of calculation work.

Let A be any constant or assumed mean. Now, let us define a new variable $d_i = x_i - A$, i = 1, 2, . . ., n. The formula for mean of X, i.e. $\bar{X}$, in terms of d_i for an individual series of X is given as,

$$\bar{X} = A + \frac{\sum_{i=1}^{n} d_i}{n} = A + \frac{\Sigma d}{N}$$

Where, n = number of observations

A = assumed mean

Again, for a grouped frequency distribution the formula is given as,

$$\bar{X} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} = A + \frac{\Sigma fd}{N}$$

Where, N = total frequency = $\sum_{i=1}^{n} f_i$

A = assumed mean

Assuming the observation x_i occurs f_i times, i = 1, 2, . . ., n and $d_i = x_i - A$, i = 1, 2, . . ., n

Step Deviation Method or Coding Method

In a grouped frequency distribution, if all the classes are of equal width, say h, the successive mid-values of various classes will differ from each other by this width. This fact can be utilised for reducing the work of computations.

Let us define $d_i' = \frac{x_i - A}{h}$, i = 1, 2, . . ., n

Where, A = assumed mean

h = class width

The formula for mean of X, i.e. $\bar{X}$, in terms of d_i' is given as,

$$\bar{X} = A + \frac{\sum_{i=1}^{n} f_i d_i'}{N} \times h = A + \frac{\Sigma fd'}{N} \times h$$

Example 1: Following are the daily wages in rupees of a sample of 9 workers: 58, 62, 48, 53, 70, 52, 60, 84, 75. Compute the mean wage.

Solution : Let x denote the daily wage in rupees.

Then as given, x_1 = 58, x_2 = 62, x_3 = 48, x_4 = 53, x_5 = 70, x_6 = 52, x_7 = 60, x_8 = 84 and x_9 = 75.

Thus, the mean wage, i.e. $\bar{X}$, is given by,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$= \frac{1}{9} \times (58+62+48+53+70+52+60+84+75)$$

$$= \frac{1}{9} \times 562$$

= Rs. 62.44

Example 2: Compute the mean weight of a group of students of a college from the following data :

Weight in kgs. :	44-48	49-53	54-58	59-63	64-68	69-73
No. of students :	3	4	5	7	9	8

Solution : Let us prepare the following table :

(1) Weight in kgs	(2) No. of students (f_i)	(3) Mid-Value (x_i)	(4) = (2) x (3) $f_i x_i$
44-48	3	46	138
49-53	4	51	204
54-58	5	56	280
59-63	7	61	427
64-68	9	66	594
69-73	8	71	568
Total = N = 36			$\Sigma fi\, xi = 2211$

∴ Mean weight of 36 students is given as,

$$\bar{X} = \frac{1}{n}\Sigma f_i x_i$$

$$= \frac{1}{36} \times 2211 \text{ kgs.}$$

= 61.42 kgs.

Example 3: The following is the frequency distribution of age of 670 students of a school. Compute the arithmetic mean of the data.

Age (in years) :	5	6	7	8	9	10	11	12	13	14
No. of students :	25	45	90	165	112	96	81	26	18	12

Solution : To apply Assumed Mean formula, let us prepare the following table taking A = 8 as assumed mean in the formula.

x (Age)	f (No. of student)	d = x = A = x – 8	fd
5	25	–3	–75
6	45	–2	–90
7	90	–1	–90

8	165	0	0
9	112	1	112
10	96	2	192
11	81	3	243
12	26	4	104
13	18	5	90
14	12	6	72
	N = 670		$\Sigma fd = 558$

$\therefore$ AM of age of the 670 students is given as,

$$\bar{x} = A + \frac{\Sigma fd}{N} = 8 + \frac{558}{670}$$

= 8 + 0.83 = 8.83 years

Example 4: Find the AM for the following distribution :

Class Interval :	350-369	370-389	390-409	410-429
Frequency :	23	38	58	82
Class Interval :	430-449	450-469	470-489	
Frequency :	65	31	11	

Solution : To find the AM of the distribution, let us use the Step-Deviation Method to simplify our calculations. Now, we prepare the following table:

Class Interval	Frequency (f)	Mid-value (x)	$d' = \frac{x-A}{h} = \frac{x-419.50}{20}$	fd'
350-369	23	359.50	–3	–69
370-389	38	379.50	–2	–76
390-409	58	399.50	–1	–58
410-429	82	419.50 (A)	0	0
430-449	65	439.50	1	65
450-469	31	459.50	2	62
470-489	11	479.50	3	33
	N = 308			$\Sigma fd' = -43$

$\therefore$ The required AM is given by,

$$\bar{X} = A + \frac{\Sigma fd'}{N} \times h$$

$= 419.50 + \frac{(-43)}{308} = 419.50 - 2.79 = 416.71$

Example 5: The following distribution gives the pattern of overtime work done by 100 employees of a company. Calculate the average overtime work done per employee.

Overtime (in hrs) :	10-15	15-20	20-25	25-30	30-35	35-40
No. of employees :	11	20	35	20	8	6

Solution: Let us prepare the following table taking A=22.5 as assumed mean in the formula.

Overtime (in hrs)	No. of employee	Mid-value (x)	$d' = \frac{x-A}{h}$ $= \frac{x-225}{5}$	$f.d'$
10-15	11	12.5	-2	-22
15-20	20	17.5	-1	-20
20-25	35	22.5 ←A	0	0
25-30	20	27.5	1	20
30-35	8	32.5	2	16
35-40	6	37.5	3	18
	N=100			$\sum fd' = 12$

∴ The required AM is given by,

$$\overline{X} = A + \frac{\sum fd'}{N} \times h$$

$$= 22.5 + \frac{12}{100} \times 5 = 23.1 \text{ hrs.}$$

Properties of Arithmetic Mean

(i) If all the observations assumed by a variable are constants, say k, then the AM is also k.

For example, if the height of every student in a group of 10 students is 170 cm, then the mean height is, of course, 170 cm.

(ii) The algebric sum of deviations of a set of observations from their AM is zero. i.e. for individual series, $\Sigma(x_i - \bar{x}) = 0$ and for frequency distribution, $\Sigma f_i(x_i - \bar{x}) = 0$

For example, if a variable X assumes five observations, say, 58, 63,

37, 45, 29, then $\bar{x}$ = 46.4. Hence, the deviations of the observations from the AM i.e. $(x_i - \bar{x})$ are (58 – 46.4) = 11.60, (63–46.4) = 16.60, (37–46.4) = –9.40, (45–46.4) = –1.40, (29–46.4) = –17.40. Now, $\Sigma(x_i - \bar{x})$ = 11.60+16.60+(–9.40)+(–1.40)+(–17.40) = 0.

(iii) AM is affected due to change of origin and/or scale which implies that if the original variable x is changed to another variable y so that, $y = \frac{x \pm a}{b}$ then $\bar{y} = \frac{\bar{x} \pm a}{b}$.

For example, suppose the AM of the values of a variable x is 25. Now, if each value of x is increased by 5, we will get a new variable defined as,

$y = x + 5$

$\therefore$ the AM of the new variable would be,

$\bar{y} = \bar{x} + 5$ = 25 + 5 = 30

i.e. after increament in each value of x, the new AM is 30.

(iv) If there are two groups containing n_1 and n_2 observations and $\bar{x}_1$ and $\bar{x}_2$ are the respective arithmetic means, then the combined AM is given by, $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$

This property could be extended to k groups and we may write

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

Example 6: The mean salary for a group of 40 female workers is Rs. 5200 per month and that for a group of 60 male workers is Rs. 6800 per month. What is the combined salary?

Solution : As given, n_1 = 40, n_2 = 60, $\bar{x}_1$ = Rs. 5200 and $\bar{x}_2$ = Rs. 6800 Hence, the combined mean salary per month is,

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$= \frac{40 \times Rs.\ 5200 + 60 \times Rs.\ 6800}{40 + 60} = \text{Rs. } 6160.$$



CHECK YOUR PROGRESS

Q 1: State true or false

(i) Wtih nominal data, the mean should be used as a measure of central tendency. (T/F)

(ii) When the date are interval or ratio, we can use the mean as a measure of central tendency. (T/F)

(iii) The mean of a data set remains unaffected if an observation equal to mean is included in it. (T/F)

(iv) The sum of deviations from mean is zero. (T/F)

1.4 MEDIAN

Median, for a given set of observations, may be defined as the middle-most value when the observations are arranged either in an ascending order or a descending order of magnitude. In other words, median is termed as a positional average which means that the value of the median is dependent upon the position of the given set of observations for which the median is wanted.

In order to determine the median of an individual series, first of all we have to arrange the observations either in ascending or in descending order of magnitude. Now, suppose we assume that there are n observations in the series. If n is odd i.e. there are odd number of observations in the series, then the $\left(\frac{n+1}{2}\right)^{th}$ value from the beginning will be the median. If n is even, i.e. there are even number of observations in the series, then the arithmetic mean of the $\frac{n}{2}^{th}$ value and the $\left(\frac{n}{2}+1\right)^{th}$ value will be the median.

Again, to determine the median of an ungrouped frequency distribution, first of all we have to construct a cumulative frequency table. If the total frequency (i.e. the total no. of observations) N is odd then the $\left(\frac{N+1}{2}\right)^{th}$ term will be the median. Otherwise, if N is even then the AM of the

$\frac{N}{2}^{th}$ and the $\left(\frac{N}{2}+1\right)^{th}$ terms would be the median.

To determine median from a grouped frequency distribution, first of all we have to construct a cumulative frequency table. Then we determine the median class from the table. The median class is that which contains the median value. If the total frequency N is odd then the median class will be that class which will contain the $\left(\frac{N+1}{2}\right)^{th}$ observation. Again, if N is even then the median class will be that class which will contain the $\frac{N}{2}^{th}$ observation. After determining the median class the particular median value is calculated by using the following formula :

$$\text{Median (Me)} = L + \frac{\frac{N}{2} - fc}{f} \times h$$

Where, L = lower limit of the median class

f = frequency of the median class

fc = cumulative frequency just before the median class

h = width of the median class

N = total frequency

Example 7: Find the medians of the following two series :

(i) 20, 15, 25, 28, 18, 16, 30

(ii) 245, 230, 265, 236, 220, 250

Solution : (i) Arranging the observations in ascending order,

we get, 15, 16, 18, 20, 25, 28, 30.

Since, no. of observations n = 7 i.e. odd, the median would be the $\left(\frac{7+1}{2}\right)^{th}$ term, i.e. the 4th observation.

Hence, median of the series = 20

(ii) Arranging the observations in ascending order,

we get, 220, 230, 236, 245, 250, 265

Here, no. of observations n = 6, i.e. even.

$\therefore$ Median will be AM of the $\left(\frac{n}{2}\right)^{th}$ and the $\left(\frac{n}{2}+1\right)^{th}$ terms, i.e. AM of the 3rd and 4th term.

Hence, median of the series $= \frac{236+245}{2} = 240.5$

Example 8: Find the median of the following distribution :

Variable (x) :	10	11	12	13	14	15	16
Frequency (f) :	8	15	25	20	12	10	5

Solution : Let as prepare the cumulative frequency table :

x	f	c.f.
10	8	8
11	15	23
12	25	48
13	20	68
14	12	80
15	10	90
16	5	95
	N = 95	

Here, N = 95, which is odd.

Hence, $\left(\frac{N}{2}+1\right)^{th}$ term would be the median,

i.e. $\left(\frac{95+1}{2}\right)^{th}$ term would be the median,

i.e. 48th term would be the median,

i.e. X = 12 is the median.

Example 9: Find the median of the following distribution :

Wages (Rs.) :	20	21	22	23	24	25	26	27	28
No. of workers :	8	10	11	16	20	25	19	9	6

Solution : Let us prepare the following cumulative frequency table :

Wages (Rs.)	No. of worker (frequency)	Cumulative frequency
20	8	8
21	10	18
22	11	29
23	16	45
24	20	65
25	25	90
26	19	109
27	9	118
28	6	124
	N = 124	

Here, N = 124, which is even.

Hence, median = AM of the $\left(\frac{N}{2}\right)^{th}$ and the $\left(\frac{N}{2}+1\right)^{th}$ terms

= AM of the $\left(\frac{124}{2}\right)^{th}$ and the $\left(\frac{124}{2}+1\right)^{th}$ terms

= AM of the 62nd and 63rd terms

= $\frac{24+24}{2}$ ($\because$ both the terms are living between c.f. 45 and 65)

= Rs. 24

Example 10: Determin median for the following distribution :

Daily wages (Rs.) :	50-55	55-60	60-65	65-70
No. of workers :	6	10	22	30
Daily wages (Rs.) :	70-75	75-80	80-85	
No. of workers :	16	12	15	

Solution : Let us prepare the following cumulative frequency table :

Daily Wages	No. of Workers (f)	Cumulative frequency (fc)
50-55	6	6
55-60	10	16
60-65	22	38
65-70	30	68
70-75	16	84

75-80	12	96
80-85	15	111
	N = 111	

Here, N = 111 which is odd

$\therefore \left(\frac{N+1}{2}\right)^{th}$ term = $\left(\frac{111+1}{2}\right)^{th}$ term = 56th term

From the cumulative frequency table we find that the 56th term lies in the class 65-70.

$\therefore$ 65-70 is the median class.

Now, median $= L + \frac{\frac{N}{2} - fc}{f} \times h$

Here, $L = 65, f = 30, fc = 38, N = 111, h = 5$

$\therefore$ Median $= 65 + \frac{\frac{111}{2} - 38}{30} \times 5$

$= 65 + \frac{55.5 - 38}{30} \times 5$

$= 65 + \frac{17.5}{6}$

= 65 + 2.92

= 67.92

i.e. the required median = Rs. 67.92

Example 11: Calculate median of the following data :

Weight in inches : 3-4 4-5 5-6 6-7 7-8 8-9 9-10 10-11

No. of saplings : 3 7 12 16 22 20 13 7

Solution : Let us prepare the following table :

Class Intervals	Frequency (f)	Cumulative frequency (fc)
3-5	3	3
4-5	7	10
5-6	12	22
6-7	16	38
7-8	22	60

8-9	20	80
9-10	13	93
10-11	7	100

N =100

Here, N = 100 which is even.

$\therefore \frac{N^{th}}{2}$ term = $\frac{100^{th}}{2}$ term = 50th term.

From the cumulative frequency table we find that the 50th term lies in the class 7-8.

$\therefore$ 7-8 is the median class.

Now, median $= L + \frac{\frac{N}{2} - fc}{f} \times h$

Here, L =7, f = 22, fc = 38, N = 100, h = 1.

$\therefore$ Median $= 7 + \frac{\frac{100}{2} - 38}{22} \times 1$

$= 7 + \frac{50 - 38}{22} \times 1$

$= 7 + \frac{12}{22}$

= 7 + 0.55

= 7.55 inches



CHECK YOUR PROGRESS

Q 2: Fill in the blanks :

(i) Sum of obsolute deviations from median is _____.

(ii) The median is less affected than mean by _____ values of observations in a distribution.

(iii) The average which divides a distribution into two equal parts is _____.

1.5 MODE

For a given set of observations, mode may be defined as the value that occurs maximum number of times. Thus, mode is that value which has the maximum concentration of the observations around it. This can also be deseribed as the most common value with which, even a layman may be familiar with.

To determine mode from an individual series, we identify the observation which appears maximum number of times. If more than one observation appears maximum number of times, all those observations are termed as mode. But, if all the observations of the series appears equal number of times, we say that, there is no mode of the given series.

To obtain mode for a grouped frequency distribution, first of all we have to identify the modal class. The class against which the frequency is maximum, is termed as modal class. Further, to find exact value of mode we apply the following formula :

$$\text{Mode (Mo)} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where, L = lower limit of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class.

h = width of the modal class.

Example 12: Determine mode/modes of the following series, if any :

(i) 3, 4, 5, 2, 3, 4, 1, 6, 4

(ii) 7, 9 11, 7, 6, 5, 9, 13

(iii) 3, 5, 6, 7, 9, 12, 3, 6, 5, 9, 12, 7

Solution : (i) The number 4 appears maximum number of times. Hence, 4 is the mode of the given series.

(ii) The numbers 7 and 9 appears maximum number of times i.e. 2 times. Hence, 7 and 9 both are modes of the given series.

(iii) Since, all the numbers appear equal number of times i.e. 2 times, hence the given series has no mode.

Example 13: Compute mode for the following distribution :

Class Interval :	350-369	370-389	390-409	410-429
Frequency :	23	38	58	82
Class Interval :	430-449	450-469	470-489	
Frequency :	65	31	11	

Solution : Since the frequency of the class 410-429 is the maximum, hence this class is the modal class. The continuous form of this modal class would be 409.5 – 429.5.

Now, mode $= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

Here, L = 409.5, f_1 = 82, f_0 = 58, f_2 = 65, h = 20

$\therefore$ Mode $= 409.5 + \frac{82 - 58}{2 \times 82 - 58 - 65} \times 20$

= 421.21

Example 14: The following table gives the incomplete income distribution of 300 workers of a firm, where the frequencies of the classes 3000-4000 and 5000-6000 are missing. If the mode of the distribution is Rs. 4428.57 find the missing frequencies.

Monthly Income (Rs.)	No. of Workers
1000-2000	30
2000-3000	35
3000-4000	?
4000-5000	75
5000-6000	?
6000-7000	30
7000-8000	15

Solution : Let the frequency of the class 3000-4000 be fi. Then the frequency of the class 5000-6000 will be equal to 300-30-35-fi-75-30-15=115-fi. It is given that, mode=4428.57, therefore, modal class is 4000-5000.

Thus, L=4000, h=1000

$= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$

$$\Rightarrow 4428.57 = 4000 + \frac{75 - fi}{2 \times 75 - fi - (115 - fi)}$$

$\Rightarrow fi = 60$ (after simplification)

$\therefore$ frequency of the class 3000-4000 is 60 and the frequency of the class 5000-6000 is (115-fi=115-60)=55.



CHECK YOUR PROGRESS

Q 3: Fill in the blanks :

(i) The mode is always found at the _____ point of a graph of frequency distribution.

(ii) Most appropriate average to determine the size of oranges on a tree is _____

(iii) Mode _____ based on all the observations.

(iv) Like mean or median, mode is not affected by _____ observations.

Relationship among Mean, Median and Mode

Prof. Karl Pearson has established a relationship empirically i.e. experimentally among mean, median and mode. Please note, for any symmetcial distribution, Mean=Median=Mode. However, for asymmetrical distribution, the following relationship among mean, median and mode approximately holds for:

Mean – Mode = 3 (Mean – Median)

1.6 GEOMETRIC MEAN

The geometric mean (GM) is a type of mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). If a variable X assumes n values $x_1, x_2, x_3, ..., x_n$, all the values being positive, then the Geometric Mean (GM) of X, denoted by G, is given by

$$G = (x_1 \times x_2 \times x_3 \times \ldots \times x_n)^{\frac{1}{n}}$$

Again, if the frequencies of the values $x_1, x_2, x_3, ..., x_n$ of X are $f_1, f_2, f_3, ..., f_n$ respectively then the geometric mean G of X is given by

$$G = \left(x_1^{f_1} \times x_2^{f_2} \times x_3^{f_3} \times \ldots \times x_n^{f_n}\right)^{\frac{1}{N}}$$

Where $N = \Sigma f$ = total frequency

Example 15: Find the GM of 2, 6 and 12

Solution : As given $x_1 = 2$, $x_2 = 6$, $x_3 = 12$ and n = 3

$$\therefore GM = (x_1 \times x_2 \times x_3)^{\frac{1}{n}}$$

$$= (2 \times 6 \times 12)^{\frac{1}{3}}$$

$$= (6^3)^{\frac{1}{3}}$$

= 6

Example 16: Find the GM for the following distribution :

x :	2	4	8	16
y :	2	3	3	2

Solution : Here, N = total frequency $= \Sigma f = 10$

$$\therefore GM = \left({x_1}^{f^1} \times {x_2}^{f2} \times {x_3}^{f3} \times {x_4}^{f^n}\right)^{\frac{1}{N}}$$

$$= \left(2^2 \times 4^3 \times 8^3 \times 16^2\right)^{\frac{1}{10}}$$

$$= (2)^{2.50}$$

$$= 4\sqrt{2}$$

= 5.66

1.7 HARMONIC MEAN

The reciprocal of the arithmetic mean of the reciprocals of the values of a variable is called the harmonic mean (HM). So, if a variable x assumes n non-zero values $x_1, x_2, x_3, ..., x_n$ then the HM of x denoted by H is given by

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}} = \frac{n}{\Sigma \frac{1}{x}}$$

Again, if the frequencies of x are $f_1, f_2, f_3, ..., f_n$ then the harmonic mean of x is given by

$$H = \frac{f_1 + f_2 + f_3 + \ldots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \ldots + \frac{f_n}{x_n}} = \frac{N}{\Sigma \frac{f}{x}}$$, where $N = \Sigma f$

Example 16: Find the HM of 4, 6 and 10

Solution : Here, x_1 = 4, x_2 = 6, x_3 = 10 and n = 3

$$\therefore HM = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}} = \frac{3}{\frac{1}{4}+\frac{1}{6}+\frac{1}{10}}$$

$$= \frac{3}{0.25+0.17+0.10} = 6.77$$

Example 17: Find the HM for the following data

x :	2	4	8	16
f :	2	3	3	2

Solution : Here, N = total frequency = (Σf) = 10

$$\therefore HM = \frac{N}{\frac{f_1}{x_1}+\frac{f_2}{x_2}+\frac{f_3}{x_3}+\frac{f_3}{x_n}}$$

$$= \frac{10}{\frac{2}{2}+\frac{3}{4}+\frac{3}{8}+\frac{2}{16}}$$

= 4.44

Example 18: An investor buys Rs. 20,000 worth of shares of a company each month. During the first 3 months he bought the shares at a price of Rs. 120, Rs. 160 and Rs. 210. After 3 months what is the average price paid by him for the shares?

Solution : Since the value of shares are changing after every one month, therefore the required average price per share is the harmonic mean of the prices paid in first three months.

$$\therefore H.M = \frac{3}{\frac{1}{120}+\frac{1}{160}+\frac{1}{210}} = \frac{3}{0.008+0.006+0.004}$$

$$= \frac{3}{0.018} = \text{Rs. } 166.66$$



CHECK YOUR PROGRESS

Q 4: Fill in the blanks :

(i) If an deservation in the data set is zero, then its geometric mean is_____.

(ii) _____ is most suitable for averaging ratios and exponential

rates of changes.

(iii) _____ is the reciprocal of the A.M. of reciprocal of observations.

Relation among AM, GM and HM

(a) For any finite number of positive values of a variable, its $AM \geq GM \geq HM$

(b) For any two positive numbers, AM x HM = $(GM)^2$

1.8 WEIGHTED AVERAGE

The numbers which represent the relative importance of the quantities of a series are called the weights of these quantities. Weights may be actual or assumed. The average obtained giving due importance to the quantities of a series or distribution is called the weighted average of the series.

Let the weights attached to the quantities $x_1, x_2, x_3, ..., x_n$ be $w_1, w_2, w_3, ..., w_n$ respectively. The weighted arithmetic mean of these quantities, denoted by $\bar{x}_w$, is given by

Withted $AM = \bar{x}_w = \dfrac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{w_1 + w_2 + w_3 + \ldots + w_n}$

$$= \frac{\Sigma wx}{\Sigma w}$$

Similarly, the geometric mean of the quantities is given by

Weighted $GM = Anti\log\left(\dfrac{\Sigma w \log x}{\Sigma w}\right)$

Again, the harmonic mean of the same quantities is given by

Weighted $HM = \dfrac{\Sigma w}{\Sigma\left(\dfrac{w}{x}\right)}$

Example 19: Calculate the simple and the weighted arithmetic mean prices per quintal of coal purchased by a factory during the first six months of a year from the following data. When will the two averages be equal?

Month :	January	February	March	April	May	June
Price per quintal (Rs.) :	212.50	256.25	250	260	221.25	270
Quantity purchased (in quintal) :	25	30	40	50	10	45

Solution : Simple AM = Rs. $\left(\frac{212.50 + 256.25 + \ldots + 270}{6}\right)$

= Rs. $\frac{1470}{6}$

= Rs. 245.00

i.e the simple AM of price per quintal of coal = Rs. 245.00

In calculating weighted mean price, the quantities of coal purchased are to be taken as the weights of various prices.

$\therefore$ Weighted AM = Rs. $\left[\frac{(25 \times 212.50) + (30 \times 256.25) + \ldots + (45 \times 270)}{25 + 30 + \ldots + 45}\right]$

= Rs. $\frac{50362.50}{200}$

= Rs. 251.81

1.9 EXERCISES

Q 1: What is measures of central tendency? Explain the different methods of measuring central tendency.

Q 2: What are the characteristics of an ideal average? How far these are satisfied by the different methods?

Q 3: Define weighted average. Distinguish between a simple average and a weighted average.

Q 4: Complute arithmetic mean of the following data :

Marks :	0-10	10-20	20-30	30-40	40-50	50-60
No. of students :	12	18	27	20	17	6

Q 5: Calculate mean and mode from the following data :

Class Interval :	350-369	370-389	390-409	410-429
Frequency :	15	27	31	19

Class Interval :	430-449	450-469
Frequency :	13	6

(Ans. 400.58 and 394.50)

Q 6: An aeroplance flies from A to B at the rate of 500 km/hour and comes back from B to A at the rate of 700 km/hour. Find the average speed of the aeroplane.

(Ans. 583.33 km/hour)

Q 7: Following is an incomplete distribution having modal mark as 44.

Marks :	0–20	20–40	40–60	60–80	80–100
No. of students :	5	18	?	12	5

Find the missing value. (Ans. 48)

Q 8: Multiple choice questions :

(i) If the AM and GM for two numbers are 6.50 and 6 respectively then the two numbers are

(a) 6 and 7 (b) 9 and 4

(c) 10 and 3 (d) 8 and 5

[Ans. (b)]

(ii) If the AM and HM for two numbers are 5 and 3.2 respectively then the GM will be

(a) 16.00 (b) 4.10

(c) 4.05 (d) 4.00

[Ans. (d)]

(iii) For open-end classification, which of the following is the best measure of central tendency?

(a) AM (b) GM

(c) Median (d) Mode

[Ans. (c)]

(iv) When a firm registers both profits and losses, which of the following measure of central tendency cannot be considered?

(a) AM (b) GM

(c) Median (d) Mode

[Ans. (b)]

(v) If a constant 5 is added to each value of a set, the AM of the new

set is

(a) increased by 5 (b) decreased by 5

(c) not affected (d) ten times the original AM

[Ans. (a)]

(vi) If each observation of a set is divided by 2, then the mean of the new values

(a) is same as the original mean (b) is decreased by 2

(b) is half of the original mean (d) remaining the same

[Ans. (c)]

(vii) The geometric mean of $\frac{1}{32}$ and $\frac{8}{25}$ is

(a) 5 (b) 15

(c) 10 (d) 100

[Ans. (c)]

(viii) Which one of the following is not uniquely defined?

(a) Mean (b) Median

(c) Mode (d) All of these measures

[Ans. (c)]

(ix) What is the median for the following observations?

5, 8, 6, 9, 11, 4

(a) 6 (b) 7

(c) 8 (d) None of these

[Ans. (b)]

(x) If the AM and GM for 10 observations are both 15, then the value of HM is

(a) Less than 15 (b) More than 15

(c) 15 (d) cannot be datermined

[Ans. (c)]



1.10 LET US SUM UP

- Central tendency may be difined as the tendency of a given set of observations to cluster around a single central or middle value and the

single value that represents the given set of observations is described as a measure of central tendency or average.

- Following are the different measures of central tendency :
  (i) Arithmetic Mean (AM) or Mean
  (ii) Median (Me)
  (iii) Mode (Mo)
  (iv) Geometric Mean (GM)
  (v) Harmonic Mean (HM)
- For a given set of observations, the Arithmatic Mean (AM) may be defined as the sum of all the observations to be divided by the number of observations.
- Median, for a given set of observations, may be defined as the middle-most value when the observations are arranged either in an ascending order or a descending order of magnitude.
- Mode is that value which has the maximum concentration of the observations around it.
- Prof. Karl Pearson has established a relationship empirically i.e. experimentally among mean, median and mode. For any distribution the following relationship among mean, median and mode approximately holds:

  Mean – Mode = 3 (Mean – Median)
- The geometric mean is a type of mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum).
- The reciprocal of the arithmetic mean of the reciprocals of the values of a variable is called the harmonic mean.
- Relation among AM, GM and HM
  (a) For any finite number of positive values of a variable, its $AM \geq GM \geq HM$
  (b) For any two positive numbers, AM x HM = $(GM)^2$



1.11 FURTHER READING

1. A.M. Gun, M.K. Gupta, B. Dasgupta, "Fundamentals Of Statistics", Volume two, Seventh Edition.
2. Richard I. Levin, David S. Rubin,"Statistics For Management", Seventh Edition.



1.12 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q. No. 1: (i) F (ii) T (iii) T (iv) T

Ans. to Q. No. 2: (i) minimum (ii) Extreme (iii) median

Ans. to Q. No. 3: (i) highest (ii) mode (iii) isn't (iv) extreme

Ans. to Q. No. 4: (i) zero (ii) geometric mean

(iii) Harmonic Mean



1.13 MODEL QUESTIONS

A. SHORT TYPE QUESTIONS

Q 1: Write down a few properties of Arithmatic Mean.

Q 2: Show the relationship between Mean, Median and Mode.

Q 3: State the relationship between Arithmetic Mean, Geometric Mean and Harmonic Mean.

B. LONG QUESTIONS

Q 1: Discuss the step deviation method of calculating Arithmetic Mean with suitable example.

Q 2: Determine mean, median and mode from the following statistical distribution :

Class Interval :	6.5–7.5	7.5–8.5	8.5–9.5	9.5–10.5
Frequency :	5	12	25	48

Class Interval :	10.5–11.5	11.5–12.5	12.5–13.5
Frequency :	32	6	1

Q 3: What is the value of mean and median for the following data :

Marks :	5-14	15-24	25-34	35-44	45-54	55-64
No. of Students :	10	18	32	26	14	10

*** ***** ***

UNIT 2 : MEASURES OF DISPERSION

UNIT STRUCTURE

2.1 Learning Objectives

2.2 Introduction

2.3 Range

2.4 Quartile Deviation

2.5 Mean Deviation

2.6 Standard Deviation

2.7 Coefficient of Variation

2.8 Multiple Choice Questions

2.9 Exercises

2.10 Let Us Sum Up

2.11 Further Reading

2.12 Answers to Check Your Progress

2.13 Model Questions

2.1 LEARNING OBJECTIVES

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

- discuss the concept of dispersion and discuss its various measures
- use the computational techniques of different measures of dispersion
- write the concept of Lorenz curve and Gini coefficient.

2.2 INTRODUCTION

A measure of central tendency summarises the distribution of a variable into a single figure which can be regarded as its representative. This measure alone, however, is not sufficient to describe a distribution because there may be a situation where two or more different distributions have the same central value. Conversely, it is possible that the pattern of distribution in two or more situations is same but the values of their central tendency are different. Hence, it is necessary to define some additional summary measures to adequately represent the characteristics of a

distribution. One such measure is known as the measure of dispersion or the measure of variation.

The concept of dispersion is related to the extent of scatter or variability in observations. In short, dispersion for a given set of observations may be defined as the amount of deviation of the observations, usually, from an appropriate measure of central tendency. Measures of dispersion may be broadly classified into

1. Absolute measures of dispersion
2. Relative measures of dispersion

Absolute measures of dispersion are classified into

(i) Range (ii) Quartile Deviation (iii) Mean Deviation

(iv) Standard Deviation.

Likewise, we have the following relative measures of dispersion :

(i) Coefficient of Range

(ii) Coefficient of Quartile Deviation

(iii) Coefficient of Mean Deviation

(iv) Coefficient of Variation.

Characteristics of a Good Measure of Dispersion

A good measure of dispersion should posses the following characteristies :

(1) It should be easy to calculate.

(2) It should be easy to understand.

(3) It should be rigidly defined.

(4) It should be based on all the observations.

(5) It should be capable of further mathematical treatment.

(6) It should not be unduly affected by extreme observations.

(7) It should not be much affected by the fluctuations of sampling.

2.3 RANGE

For a given set of observations, range may be defined as a difference between the largest and smallest observation. Thus if L and S denote the largest and smallest observations respectively then we have

Range = L – S

The corresponding relative measure of dispersion, known as coefficient of range, is given by

$$\text{Coefficient of range} = \frac{L-S}{L+S} \times 100$$

Example 1: Find range and coefficient of range for each of the following data :

(a) Weekly wages of 10 workers of a factory are :

310, 350, 420, 105, 115, 290, 245, 450, 300, 375

(b) The distribution of marks obtained by 100 students,

Marks :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students :	6	14	21	20	18	10	5
Marks :	70-80	80-90	90-100				
No. of Students :	3	2	1				

(c) The age distribution of 60 school going children :

Age (in years) :	5-7	8-10	11-13	14-16	17-19
Frequency :	20	18	10	8	4

Solution :

(a) The largest and smallest wages are L = 450 and S = 105.

$\therefore$ Range = = L – S = 450 – 105 = 345

and coefficient of range $= \frac{L-S}{L+S} = \frac{450-105}{450+105} = 0.62$

(b) Here, the lowest class limit, that is, 5 = 0

and the highest class limit, that in, L = 100

$\therefore$ Range = L – S = 100 – 0 = 100

and coefficient of range $= \frac{L-S}{L+S} = \frac{100-0}{100+0} = 1$

(c) Here, the lowest class boundary, that is, S = 4.5

and the highest class boundary, that is, L = 19.5

$\therefore$ Range = L – S = 19.5 – 4.5 15

and coefficient of range $= \frac{L-S}{L+S} = \frac{19.5-4.5}{19.5+4.5} = 0.625$

2.4 QUARTILE DEVIATION

Half of the deviation between the third quartile (Q_3) and the first quartile (Q_1) of a distribution is called the quartile deviation (Q. D.) of the distribution. Thus,

Quantile Deviation (Q. D.) = $\frac{Q_3 - Q_1}{2}$

and coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$

Example 2: For a distribution, the values of the first quartile and the third quartile are 39.25 and 80.25. Find the quartile deviation and coefficient of quartile deviation.

Solution: We are given, Q_1 = 39.25 and Q_3 = 80.25

$\therefore$ Quantile Deviation (Q. D.) = $\frac{Q_3 - Q_1}{2}$

$$= \frac{80.25 - 39.25}{2} = \frac{41}{2} = 20.50$$

Example 3: Compute the quantite deviation and co-efficient of quartile deviation from the following distribution :

Weekly wages (in Rs.)	No. of Workers
0-100	12
100-200	18
200-300	35
300-400	42
400-500	50
500-600	45
600-700	20
700-800	8

Solution : Let us prepare the following table :

C.I.	Frequency	Comulative frequency
0-100	12	12
100-200	18	30
200-300	35	65
300-400	42	107
400-500	50	157
500-600	45	202
600-700	20	222
700-800	8	230
	N=230	

Calculation of Q_1

Since $\frac{N}{4}=\frac{230}{4}=57.5$, therefore Q_1 in the class interval 200-300.

$\therefore L_1 =$ lower limit of the Q_1 class = 200

$f_1 =$ frequency of the Q_1 class = 35

$h_1 =$ width of the Q_1 class = 100

$f_{c_1} =$ cumulative frequency just before Q_1 class=30

Hence, $Q_1 = L_1 + \frac{\frac{N}{4} - f_{c_1}}{f_1} \times h_1$

$$= 200 + \frac{57.5 - 30}{35} \times 100 = \text{Rs. } 278.57$$

Calculation of Q_3

Since $\frac{3N}{4}=\frac{3\times 230}{4}=172.5$, therefore Q_3 lies in the class interval 500-600.

$\therefore L_3 =$ lower limit of the Q_3 class = 500

$f_3 =$ frequency of the Q_3 class = 45

$h_3 =$ width of the Q_3 class = 100

$f_{c_3} =$ cumulative frequency just before Q_3 class=157

Hence, $Q_3 = L_3 + \frac{\frac{3N}{4} f_{c_3}}{f_3} \times h_3$

$$= 500 + \frac{172.5 - 157}{45} \times 100$$

= Rs. 534.44

$$\therefore Q.D. = \frac{Q_3 - Q_1}{2} = \frac{534.44 - 278.57}{2} = \text{Rs } 127.94$$

and coefficient $Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{534.44 - 278.57}{534.44 + 278.57} = 0.315$



CHECK YOUR PROGRESS

Q 1: Fill in the blanks :

(i) Range is a measure of dispersion most affected by _____ observations.

(ii) An appropriate measure of dispersion with open end class interval is_____.

(iii) Quartile deviation takes into account_____ 50% of the observations.

2.5 MEAN DEVIATION

For a given set of observation, mean deviation is defined as the arithmetic mean of the absolute deviation of the observations from an appropriate measure of central tendency. Hence, if a variable X assumes n values $x_1, x_2, ..., x_n$, then the mean deviation of X about an average A is given by,

$$MD_A = \frac{1}{n}\sum_{i=1}^{n}|x_i - A| = \frac{1}{n}\sum|x - A|$$

For a grouped frequency diftribution, mean deviation about an average A is given by

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

$$\frac{1}{N}\sum f|x - A|$$

Where x_i and f_i denote the mid value and frequency of the ith-class interval and

$N = \sum f_i$ = total frequency

A relative measure of dispersion applying mean deviation is given by

Coefficient of mean deviation = $\frac{Mean\ deviation\ about\ A}{A} \times 100$

Example 4: What is the mean deviation about mean for the following numbers?

5, 8, 10, 10, 12, 9

Solution : The mean is given by

$$\bar{x} = \frac{5+8+10+10+12+9}{6}$$

= 9

Now we prepare the following table :

x	$\lvert x-\bar{x}\rvert$
5	4
8	1
10	1
10	1
12	3
9	0
	Total = $\Sigma\lvert x-\bar{x}\rvert=10$

Thus mean deviation about mean is given by

$$M.D.=\frac{1}{n}\Sigma\lvert x-\bar{x}\rvert=\frac{10}{6}=1.67$$

Example 5: Compute the mean deviation about the arithmetic mean for the following date :

X :	1	3	5	7	9
f :	5	8	9	2	1

Also find the coefficient of the mean deviation about the AM.

Solution : Let us prepare the following table :

x	f	fx	$\lvert x-\bar{x}\rvert=\lvert x-3.88\rvert$	$f\lvert x-\bar{x}\rvert$
1	5	5	2.88	14.40
3	8	24	0.88	7.04
5	9	45	1.12	10.08
7	2	14	3.12	6.24
9	1	9	5.12	5.12
	$N=\Sigma f=25$	$\Sigma fx=97$		$\Sigma f\lvert x-\bar{x}\rvert=42.88$

Here, $AM=\bar{x}=\frac{1}{N}\Sigma fx$

$$=\frac{1}{25}\times 97=3.88$$

$\therefore$ Mean deviation about AM = $=\frac{1}{N}\sum f\left|x-\bar{x}\right|=\frac{1}{25}\times 42.88$

Also the coefficient of MD about its AM is $=1.72$

$$\frac{MD\ about\ AM}{AM}\times 100$$

$$= \frac{1.72}{3.88} \times 100$$

= 44.33

Example 6: Compute the coefficient of mean deviation about median for the following distribution :

Weight in kgs :	40-50	50-60	60-70	70-80
No. of persons :	8	12	20	10

Solution : We need to compute the median weight in the first stage.

Let us prepare the following table :

Wt. (in kgs)	No. of persons (f)	Cumulative trequency (fi)	Mid value (x)	$\lvert x - Me \rvert = \lvert x - 62.5 \rvert$	$f \lvert x - Me \rvert$
40-50	8	8	45	17.5	140
50-60	12	29	55	7.5	90
60-70	20	40	65	2.5	50
70-80	10	50	75	12.5	125
	N = 50				$\Sigma f \lvert x - Me \rvert = 405$

Here, $\frac{N^{th}}{2}$ term = $\frac{50^{th}}{2}$ term = 25th term. From the table we find that this term is included in the class 60-70. Hence this class is the median class.

$\therefore$ Median, $Me = L + \frac{\frac{N}{2} - fc}{f} \times h$

$$= 60 + \frac{25 - 20}{20} \times 10$$

= 62.50 kg.

Thus, mean deviation about median

$$= \frac{1}{N} \Sigma f \lvert x - Me \rvert = \frac{1}{50} \times 405\, kg$$

= 8.10 kg.

Hence, coefficient of mean deviation about median

$$= \frac{M.D.\ about\ median}{Median} \times 100$$

$$= \frac{8.10}{62.50} \times 100$$

= 12.96

2.6 STANDARD DEVIATION

Although mean deviation is an improvement over range so far as a measure of dispersion is concerned, mean deviation is difficult to compute is further more, it cannot be treated mathematically. The best measure of dispersion is, usually, standard deviation which does not possess the demerits of range and mean deviation.

Standard Deviation for a given set of observations is defined as the root mean square deviation when the deviations are taken from the AM of the observations. If a variable x assumes n values x_1, x_2, x_3, ..., x_n then its standard diviation, generally denoted by σ, is given by

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

$$= \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} \qquad \text{(simplified form)}$$

For a grouped frequency distribution, the standard deviation is given by

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

$$= \sqrt{\frac{1}{N}\Sigma f(x-\bar{x})^2}$$

$$= \sqrt{\frac{1}{N}\Sigma fx^2 - \bar{x}^2} \qquad \text{(simplified form)}$$

Assumed Mean Method

In case of individual series we have the following formula :

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}, \qquad \text{d = x–A where A = assumed mean}$$

In case of frequency distribution,

$$\sigma = \sqrt{\frac{fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

Step Deviation Method

If the deviations from assumed mean A have some common factors then these deviations are divided by the highest common factor h (say) and the step deviations denoted by d' are obtained.

Thus $d' = \frac{x - A}{h}$ where h is the H.C.F. of the deviations d = x = A. In this method, we have the following formula :

For individual series,

$$\sigma = \sqrt{\frac{fd'^2}{n} - \left(\frac{\Sigma d'}{n}\right)^2} \times h$$

For frequency distributions,

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$$

where $d' = \frac{d}{h} = \frac{x - A}{h}$, A = assumed mean

Empirisal Relationship among Q.D., M.D. and S.D.

It can be verified empirically that for a distribution the following relationship among Q. D., M. D. and S. D. approximately holds. However, in case of normal distribution, this relationship is exact.

$$Q.D. = \frac{5}{6} M.D. = \frac{2}{3} S.D.$$

Or 6 Q. D. = 5 M. D. = 4 S. D.

Coefficient of Variation

The standard deviation is an absolute measure of dispersion and is expressed in the same units as the units of variable X. A relative measure of dispersion, based on standard deviation is known as coefficient of standard deviation or coefficient of variation and is given by

$$C.V. = \frac{\sigma}{\bar{x}} \times 100$$

This measure introduced by karl pearson, is used to compare the

variability or homogeneity or stability or uniformity or consistency of two or more sets of data having different units of measurement. The data having a higher value of the coefficient of variation is said to be more dispersed or less uniform, etc.

Combined or Composite Standard Deviation

Let a distribution having n_1 observations has mean $\bar{x}_1$ and standard divition σ_1 and let another distribution having n_2 observations has mean $\bar{x}_2$ and standard deviation σ_2. Then the standard deviation σ of the distribution obtained by combining the two distributions having altogether $(n_1 + n_2)$ observations is given by

$$\sigma = \sqrt{\frac{n_1\sigma_1^{\,2} + n_2\sigma_2^{\,2} + n_1d_1^{\,2} + n_2d_2^{\,2}}{n_1 + n_2}}$$

Where $d_1 = \bar{x}_1 - \bar{x}, \quad d_2 = \bar{x}_2 = \bar{x}$

and $\bar{x}$ = combined mean = $\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

Variance : The square of standard deviation is called variance. Conversely, the positive square root of variance is standard deviation, since standard deviation is by definition non-negative. If standard deviation of a variable X in given as σ_x or σ, then the variance of X is given as

Var (x) or V (x) = $\sigma_x^{\,2}$ or σ^2

Properties of Standard Deviation

1. If all the observations assumed by a variable are constant, i.e. equal, then the S. D. is zero. This means that if all the values taken by a variable x is k, say, then S. D. of x i.e. $\sigma_x = 0$. This result applies to range as well as mean deviation.
2. Standard deviation is independent of change of origin but not of change of scale. This property implies that if a constant is added (or subtracted) to all the observations, the value of standard deviation remains unaffected. On the other hand, if all the observations are multiplied or divided by a constant, the standard deviation also gets multiplied (or divided) by this constant.
3. Standard deviation of a given set of observations is not less than mean

deviation from mean, i.e., Standard Deviation $\geq$ Mean Deviation from mean.

4. Standard deviation of a given set of observations is not greater than any other root mean square deviation, i.e. $\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2} \leq \sqrt{\frac{1}{n}\Sigma(x-A)^2}$

Example 7: Find the standard deviation of the following distribution :

Weights (kgs) :	50-52	52-54	54-56	56-58	58-60
No. of students :	17	35	28	15	5

Solution : Let us prepare the following table :

Weight (kgs)	No. of students (f)	Mid Value (x)	$d' = \frac{x-A}{h} = \frac{x-55}{2}$	fd'	fd'^2
50-52	17	51	–2	–34	68
52-54	35	53	–1	–35	35
54-56	28	55	0	0	0
56-58	15	57	1	15	15
58-60	5	59	2	10	20
	$N = \Sigma f = 100$			$\Sigma fd' = -44$	$\Sigma fd'^2 = 138$

Applying step deviation method we have,

$$S.D.(\sigma) = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$$

$$= \sqrt{\frac{138}{100} - \left(\frac{-44}{100}\right)^2} \times 2 \text{ kgs.}$$

$$= \sqrt{1.38 - 0.1936} \times 2 \text{ kgs.}$$

= 2.18 kgs.

Example 8: Calculate standard deviation and coefficient of variation from the following data :

Age (year) :	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of students :	3	61	132	153	140	51	2

Solution : Let us prepare the following table :

Age	No. of students (f)	Mid Value (x)	$d' = \frac{x-A}{h} = \frac{x-55}{10}$	fd'	fd'^2
20-30	3	25	–3	–9	27
30-40	61	35	–2	–122	244
40-50	132	45	–1	–132	132
50-60	153	55	0	0	0
60-70	140	65	1	140	140
70-80	51	75	2	102	204
80-90	2	85	3	6	18
	$N = \Sigma f = 542$			$\Sigma fd' = -15$	$\Sigma fd'^2 = 765$

$$S.D.(\sigma) = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$$

$$= \sqrt{\frac{765}{542} - \left(\frac{-15}{542}\right)^2} \times 10$$

$$= \sqrt{1.4114 - 0.0009} \times 10$$

$$= \sqrt{1.4105} \times 10$$

= 11.876 years

Again, $C.V. = \frac{\sigma}{\bar{x}} \times 100$

Now, $\bar{x} = A + \frac{\Sigma fd'}{N} \times h$

$$= 55 + \frac{-15}{542} \times 10$$

= 55 – 0.277 = 54.723 years

$$\therefore\ C.V. = \frac{11.876}{54.723} \times 100 = 21.4$$

Example 9: For a group of 60 boy students, the mean and SD of statistics marks are 45 and 2 respectively. The same figures for a group of 40 girl students are 55 and 3 respectively. What is the mean and SD of marks if the two groups are pooled together?

Solution : As given, $x_1 = 60,\ \bar{x}_1 = 45,\ \sigma_1 = 2$

$n_2 = 40,\ \bar{x}_2 = 55,\ \sigma_2 = 3$

Thus the combined mean is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{60\times 45 + 40\times 55}{60 + 40}$$

= 49

Thus, $d_1 = \bar{x}_1 - \bar{x} = 45 - 49 = -4$

$d_2 = \bar{x}_2 - \bar{x} = 55 - 49 = 6$

$\therefore$ combined SD is given as

$$\sigma = \sqrt{\frac{n_1 s_1^{\,2} + n_2 s_2^{\,2} + n_1 d_1^{\,2} + n_2 d_2^{\,2}}{n_1 + n_2}}$$

$$= \sqrt{\frac{60\times 2^2 + 40\times 3^2 + 60\times (-4)^2 + 40\times 6^2}{60 + 40}}$$

$= \sqrt{30}$ = 5.48

Example 10: Two teams of football players A and B gave goals as shown below. Determine with the help of C. V., the team that is more consistent.

No. of Goals	No. of Games	
	A	B
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

Solution : Let us prepare the following table :

No. of goals(x)	d=x-A = x–2	d^2	A			B		
			f_1	f_1d	f_1d^2	f_2	f_2d	f_2d^2
0	–2	4	27	–54	108	17	–34	68
1	–1	1	9	–9	9	9	–9	9
2	0	0	8	8	0	6	0	0
3	1	1	5	5	5	5	5	5
4	2	4	4	4	16	3	6	12
			$N_2 = 53$	$\Sigma f_1 d = 50$	$\Sigma f_1 d^2 = 138$	$N_2 = 40$	$\Sigma f_2 d = -32$	$\Sigma f_2 d^2 = 94$

Now,

$$\bar{x}_A = A + \frac{\Sigma f_1 d}{N_1} \qquad \bar{x}_B = A + \frac{\Sigma f_2 d}{N_2}$$

$$= 2 + \frac{-50}{53} = 1.1 \qquad = 2 + \frac{-32}{42} = 1.2$$

$$\sigma_A = \sqrt{\frac{\Sigma f_1 d^2}{N_1} - \left(\frac{\Sigma f_1 d}{N_1}\right)^2} \qquad \sigma_B = \sqrt{\frac{\Sigma f_2 d^2}{N_2} - \left(\frac{\Sigma f_2 d}{N_2}\right)^2}$$

$$= \sqrt{\frac{138}{58} - \left(\frac{-50}{53}\right)^2} \qquad = \sqrt{\frac{94}{40} - \left(\frac{-32}{40}\right)^2}$$

$$= \sqrt{2.60 - 0.81} \qquad = \sqrt{2.35 - 0.64}$$

$$= \sqrt{1.79} \qquad = \sqrt{1.71}$$

= 1.34 = 1.31

$$C.V.(A) = \frac{\sigma_A}{\bar{x}_A} \times 100 \qquad C.V.(B) = \frac{\sigma_B}{\bar{x}_B} \times 100$$

$$= \frac{1.34}{1.1} \times 100 \qquad = \frac{1.31}{1.2} \times 100$$

= 122 = 109

We have seen that C.V. (B) < C.V. (A). This implies that team A is more consistent than team B.

Example 11: Prove that for the first n natural numbers, standard deviation (SD) is $\sqrt{\frac{n^2 - 1}{12}}$

Solution : For the first n natural numbers, AM is given by,

$$\bar{x} = \frac{1 + 2 + 3 + ... + n}{n}$$

$$= \frac{1}{n}\left[\frac{n(n+1)}{2}\right] = \frac{n+1}{2}$$

$$\therefore SD = \sqrt{\frac{1}{n}\sum\left(x - \bar{x}\right)^2}$$

$$= \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2}$$

$$= \sqrt{\frac{1^2+2^2+3^2+....+n^2}{n} - \left(\frac{n+1}{2}\right)^2}$$

$$= \sqrt{\frac{n(n+1)(2n+1)}{6n} - \frac{(n+1)^2}{4}}$$

$$= \sqrt{\frac{(n+1)(4n+2-3n-3)}{12}}$$

$$= \sqrt{\frac{(n+1)(n-1)}{12}}$$

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

Example 12: If AM and coefficient of variation of x are 10 and 40 respectively, what is the variance of (15-2x)?

Solution : Let, y=15-2x

$\therefore \sigma_y = 2\sigma_x$, $\because$ SD is independent of change of origin.

Again, we are given,

C.V.=40 and $\bar{x} = 10$

Thus, C.V. $= \frac{\sigma_x}{x} \times 100$

$\Rightarrow 40 = \frac{\sigma_x}{10} \times 100$

$\Rightarrow \sigma_x = 4$

$\therefore \sigma_y = 2\sigma_x = 2 \times 4 = 8$

$\therefore$ Variance of $(15-2x)$ $= \{SD \ \ of(15-2x)\}^2$

$= \{SD \ of \ \ y\}^2$

$= \{\sigma_y\}^2$

$= 8^2$

= 64



CHECK YOUR PROGRESS

Q 2: Fill in the blanks :

(i) Standard deviation is always a non-negative number.

(ii) Variance indicates the average distance of any observation in the data set from the mean.

(iii) Standard deviation can be calculated by taking deviation from any measure of central tendency.

(iv) Mean deviation is menimum when deviations are taken from median.

2.7 COEFFICIENT OF VARIATION

Let the mean and the standard deviation of a distribution be $\bar{x}$ and σ respectively. The coefficient of variation (CV) is defined by

$$CV = \frac{\sigma}{\bar{x}} \times 100, \quad \bar{x} \neq 0$$

This measure was introduced by statistician karl pearson.

Note : 1. CV is a unitless measure of dispersion.

2. CV is the best measure to compare the variability of two distributions.

3. A distribution with less CV is considered more consistent (less variable).

4. A distribution with more CV is considered less consistent (more variable).

Example 13: The mean and standard deviation of heights of adults (age 30 years), and children (age 10 months) are given below -

	Adults	Children
Mean	160 cm	60 cm
SD	10 cm	5 cm

Show that there is a greater variation in the heights of children than that of adults.

Solution :

CV (adults) $= \frac{\sigma}{\bar{x}} \times 100 = \frac{10}{160} \times 100 = 6.25$

CV (children) $\frac{\sigma}{\bar{x}} \times 100 = \frac{5}{60} \times 100 = 8.33$

Now CV (children) > CV (adults)

$\therefore$ there is greater variation in the heights of children than that of adults.



CHECK YOUR PROGRESS

Q 3: The mean blood pressure of 25 boys is 120 mm Hg, and SD is 10 mm Hg. The mean height and SD of the same set of boys is 160 cm and 5 cm respectively. SHow that blood pressure is more variable than height.

2.8 MULTIPLE CHOICE QUESTIONS

Q 1: Which of the following statements is correct?

(a) Two distributions may have identical measures of central tendency and dispersion.

(b) Two distributions may have the identical measures of central tendency but different measures of dispersion.

(c) Two distributions may have the different measures of central tendency but identical measures of dispersion.

(d) All the statements (a), (b) and (c)

[Ans. (d)]

Q 2: If the mean diviation of a distribution is 6.4, the standard deviation is:

(a) 6.8 (b) 8.0

(c) 6.4 (d) None of the above

[Ans. (b)]

Q 3: The appropriate measure of dispersions for open end classifisation is

(a) Standard Deviation (b) Mean Deviation

(c) Quartile Deviation (d) All these measures

[Ans. (c)]

Q 4: A shift of origin has no impact on

(a) Range (b) Mean Deviation

(c) Standard Deviation (d) All these and quartile deviation

[Ans. (d)]

Q 5: The standard deviation of 10, 16, 10, 16, 10, 10, 16, 16 is

(a) 4 (b) 6 (c) 3 (d) 0

(Ans. 3)

Q 6: For any two numbers standard deviation is always

(a) Taric the range (b) Half of the range

(c) Square of the range (d) None of these

[Ans. (b)]

Q 7: If each observation of a series is divided by 5, the S. D. of the new observations is

(a) $\frac{1}{10}^{th}$ of S. D. of original observations

(b) 5 times of S. D. of original observations

(c) not changed

(d) $\frac{1}{5}^{th}$ of S. D. of original observations

[Ans. (d)]

Q 8: Which of the following is a unitless measure of dispersion?

(a) Standard Deviation (b) Mean Deviation

(c) Coefficient of Variation (d) Range

[Ans. (c)]

2.9 EXERCISES

Q 1: The mean and standard deviation of 10 observations are 35 and 2 respectively. Find the changed mean and standard deviation if,

(a) each observation is increased by 4 and

(b) each observation is multiplied by 5

[Ans. (a) 39, 2, (b) 175, 10]

Q 2: Determine mean deviation from median and standard deviation from the following distribution :

X :	10	11	12	13	14
f :	3	12	11	12	3

(Ans. 0.75, 1)

Q 3: Find the coefficient of variation for the following distribution of wages.

Daily wages (Rs.) :	30-40	40-50	50-60	60-70	70-80	80-90
No. of workers :	17	28	21	15	13	6

(Ans. 26.93)

Q 4: Determine quantile deviation and its coefficient from the following distribution.

Marks :	6	7	8	9	10	11	12
No. of students :	3	6	9	13	8	5	4

(Ans. 1, 0.11)

Q 5: Which of the following companies A and B is more consistent so far as the payment of dividend are concerned ?

Dividend paid by A :	5	9	6	12	15	10	8	10
Dividend paid by B :	4	8	7	15	18	9	6	6

(Ans. Company A)



2.10 LET US SUM UP

- A measure of central tendency summarises the distribution of a variable into a single figure which can be regarded as its representative.
- The concept of dispersion is related to the extent of scatter or variability in observations. In short, dispersion for a given set of observations may be defined as the amount of deviation of the observations, usually, from an appropriate measure of central tendency.
- Measures of dispersion may be broadly classified into
  1. Absolute measures of dispersion
  2. Relative measures of dispersion
- Absolute measures of dispersion are classified into
  (i) Range (ii) Quartile Deviation (iii) Mean Deviation
  (iv) Standard Deviation.
- Various relative measures of dispersion are Coefficient of Range, Coefficient of Quartile Deviation, Coefficient of Mean Deviation, Coefficient of Variation.
- For a given set of observations, range may be defined as a difference between the largest and smallest observation.
- Half of the deviation between the third quartile (Q_3) and the first quartile (Q_1) of a distribution is called the quartile deviation (Q. D.) of the distribution.
- For a given set of observation, mean deviation is defined as the arithmetic mean of the absolute deviation of the observations from an appropriate

measure of central tendency.

- Standard Deviation for a given set of observations is defined as the root mean square deviation when the deviations are taken from the AM of the observations.
- The square of standard deviation is called variance.
- The graphical method of measurement of dispersion was developed by a famous economic statisticion Dr. Max O. Lorenz. He used this method to measure the inequalities of income or wealth of a society.
- Gini coefficient is based on the Lorenz Curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half-square in which the curve lies.



2.11 FURTHER READING

1. A.M. Gun, M.K. Gupta, B. Dasgupta, "Fundamentals Of Statistics", Volume two, Seventh Edition.
2. Richard I. Levin, David S. Rubin,"Statistics For Management", Seventh Edition.



2.12 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) extreme (ii) Q.D. (iii) central

Ans to Q No 2: (i) True (ii) True (iii) False (iv) True

Ans to Q No 3: CV (blood pressure) $= \frac{\sigma}{\bar{x}} \times 100 = \frac{10}{120} \times 100 = 8.3$ (approx)

$$\text{CV (height)} = \frac{\sigma}{\bar{x}} \times 100 = 3.125$$

$$= \frac{5}{100} \times 100 = 3.125$$

now CV (blood pressure) > CV (height)

$\therefore$ blood pressure is more variable than height.



2.15 MODEL QUESTIONS

A. SHORT TYPE QUESTIONS

Q 1: Which measures of dispersion is not affected by the presence of extreme observations ?

(a) Range (b) Mean Deviation

(c) Standard Deviation (d) Quartile Deviation

Q 2: If the mean of a series is 10 and its coefficient of variation is 40 percent, the variance of the series is :

(a) 4 (b) 8

(c) 12 (d) 16

Q 3: Define coefficient of variation (C. V.). Write a short note on this measure.

B. LONG QUESTIONS

Q 1: What do you mean by dispersion? Mention the various measures of dispersion.

Q 2: The following distribution is the distribution of ages of 100 students selected at random from a college consisting of 3500 students. Find the standard distribution for this distribution.

Age (in yrs.) :	16-17	17-18	18-19	19-20	20-21	21-22	22-33
No. of students. :	4	14	18	28	20	12	4

(Ans. 1.47 yrs.)

Q 3: Determine the mean deviation from the mean for the following distribution :

Class Interval :	12-15	15-18	18-21	21-24	24-27	27-30
Frequency :	2	4	5	6	3	1

(Ans. 3.33)

Q 4: Coefficients of variation of two sets of observations are 58 and 69. Their standard deviations are 21.2 and 15.6 respectively. FInd their arithmetic means.

*** ***** ***

UNIT 3 : MEASUREMENT OF ECONOMIC INEQUALITY

UNIT STRUCTURE

3.1 Learning Objectives

3.2 Introduction

3.3 Pareto's Law of Income Distribution

3.4 Log normal Distribution (Concept only)

3.5 Lorenz Curve

3.6 Coefficient of Inequality : Gini Coefficient

3.7 Let Us Sum Up

3.8 Further Reading

3.9 Answers to Check Your Progress

3.11 Model Questions

3.1 LEARNING OBJECTIVES

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

- discuss Pareto's law of income distribution
- describe the concept of log normal Distribution
- write the concept of Lorenz curve and Gini coefficient.

3.2 INTRODUCTION

The measurement of economic inequality has remained a major subject of discussion in Economics. Different concepts and theories have been put forward to deliberate on this issue. A few important concepts in this area has been taken up in this unit. Among them are the Pareto's law of distribution, Lorenze curve and the Gini Coefficient. Apart from these concepts, the concept of log normal distribution has also been discussed.

3.3 PARETO'S LAW OF INCOME DISTRIBUTION

The concept of Pareto Distribution was put forward by Vilfredo Pareto. This is also referred to as the Pareto Principle or the 80-20 Rule. The Pareto

distribution is used in describing social, scientific, and geophysical phenomena in a society. Pareto created a mathematical formula in the early 20th century that described the inequalities in wealth distribution that existed in his native country of Italy.

Pareto observed that 80 percent of the country's wealth was concentrated in the hands of only 20 percent of the population. The theory is now applied in many disciplines such as incomes, productivity, populations, and other variables.

Empirical validity of the Theory: In 1906, Vilfredo Pareto introduced the concept of the Pareto Distribution when he observed that 20 percent of the pea pods were responsible for 80 percent of the peas planted in his garden. He related this phenomenon to the nature of wealth distribution in Italy, and he found that 80 percent of the country's wealth was owned by about 20 percent of its population. In terms of land ownership as well, he further observed that 80 percent of the land was owned by a handful of wealthy citizens, who comprised about 20 percent of the population.

The definition of the Pareto Distribution was later expanded in the 1940s by Dr. Joseph M. Juran, a prominent product quality guru. Juran applied the Pareto principle to quality control for business production to show that 20 percent of the production process defects are responsible for 80 percent of the problems in most products.

Pareto Distribution Formula: The formula for calculating the Pareto Distribution is as follows:

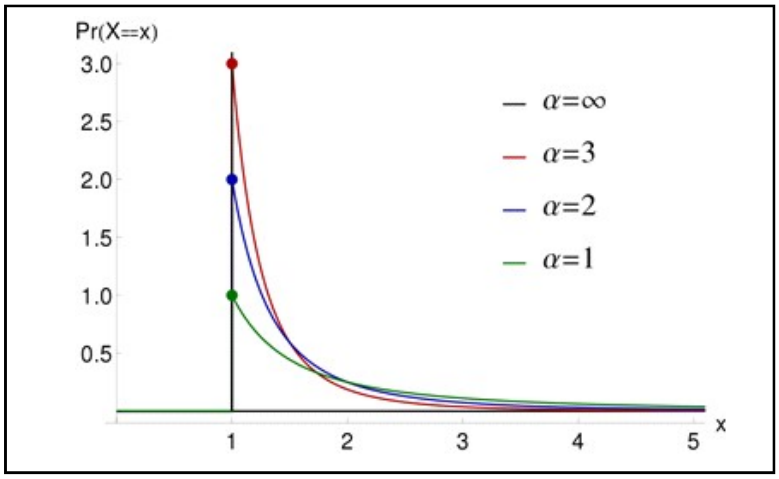
$$F(x) = 1 – (k/x)\alpha$$

Where, x – random variable, k – lower bound on data, α – shape parameter.

Graphical Representation:

On a chart, the Pareto distribution is represented by a slowly declining tail, as shown in Figure 3.1:

Figure 3.1: Pareto Distribution Chart



Source: Wikipedia Commons

The chart is defined by the variables α and x. It provides two main applications. One of the applications is to model the distribution of wealth among individuals in a country. The chart shows the extent to which a large portion of wealth in any country is owned by a small percentage of the people living in that country.

The second application is to model the distribution of city populations, where a large percentage of the population is concentrated in the urban centers and a lower amount in the rural areas. The population in urban centers continues to increase while the rural population continues to decline as younger members of the population migrate to urban centers.

Some Practical Applications of the Pareto Distribution

- **Business Management:** One of the applications of the Pareto concept is in business management. A business may observe that 20 percent of the effort dedicated to a specific business activity generates 80 percent of the business results. A business can use this ratio to identify the most important segments that it can focus on and thereby increase its efficiency.

  For example, if marketing contributed to increased business results, the business can allocate more time and resources into marketing activities to increase the company's revenues and profits.

- **Company Revenues:** The 80-20 Pareto rule may also apply in evaluating the source of the company revenues. For example, when

the company observes that 80 percent of reported annual revenues come from 20 percent of its current customers, it can focus its attention on increasing the customer satisfaction of influential customers.

From this observation, the company can also deduce that 80 percent of customer complaints come from 20 percent of customers who form the bulk of its transactions. Also, focusing on solving the complaints of 20 percent of its customers can increase the overall customer satisfaction of the company. The company should focus on retaining 20 percent of its influential customers and on acquiring new customers.

- **Employee Evaluation:** A company can also use the 80-20 rule to evaluate the performance of its employees. The company may observe that 80 percent of its overall output is the direct result of about 20 percent of its employees. Using the ratio, the company can focus on rewarding the 20 percent most productive employees as a way of motivating them and encouraging the lower cluster of employees to work harder. The productivity ratio could also show the company that 80 percent of human resource problems are caused by 20 percent of the company's employees.

Limitations of the Pareto Distribution: Some of the limitations of this principle are:

- While the 80-20 Pareto distribution rule applies to many disciplines, it does not necessarily mean that the input and output must be equal to 100 percent. For example, 20 percent of the company's customers could contribute 70 percent of the company's revenues. The ratio brings a total of 90 percent.
- Pareto concept is merely an observation that suggests that the company should focus on certain inputs more than others.

3.4 LOG NORMAL DISTRIBUTION (CONCEPT ONLY)

A log-normal distribution is a statistical distribution of logarithmic values from a related normal distribution. A log-normal distribution can be translated to a normal distribution and vice versa using associated logarithmic calculations.

A normal distributions is a probability distribution of outcomes that is symmetrical or forms a bell curve. In a normal distribution 68% of the results fall within one standard deviation and 95% fall within two standard deviations.

While most people are familiar with a normal distribution, they may not be as familiar with log-normal distribution. A normal distribution can be converted to a log-normal distribution using logarithmic mathematics. That is primarily the basis as log-normal distributions can only come from a normally distributed set of random variables.

There can be a few reasons for using log-normal distributions in conjunction with normal distributions. In general most log-normal distributions are the result of taking the natural log where the base is equal to e=2.718. However, the log-normal distribution can be scaled using a different base which affects the shape of the lognormal distribution.

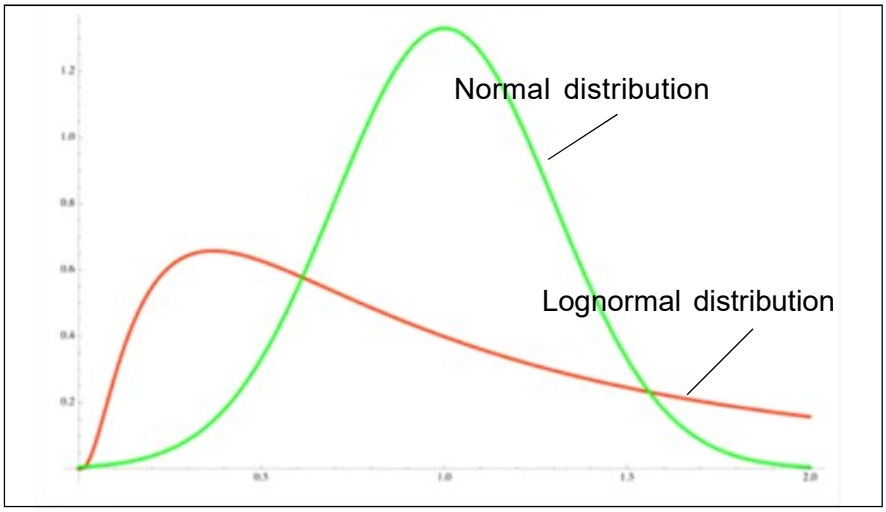
Overall the log-normal distribution plots the log of random variables from a normal distribution curve. In general, the log is known as the exponent to which a base number must be raised in order to produce the random variable (x) that is found along a normally distributed curve.

Applications and Uses of Log-Normal Distribution: Normal distributions may present a few problems that log-normal distributions can solve. Mainly, normal distributions can allow for negative random variables while log-normal distributions include all positive variables.

One of the most common applications where log-normal distributions are used in finance is in the analysis of stock prices. The potential returns of a stock can be graphed in a normal distribution. The prices of the stock however can be graphed in a log-normal distribution. The log-normal distribution curve can therefore be used to help better identify the compound return that the stock can expect to achieve over a period of time.

Note that log-normal distributions are positively skewed with long right tails due to low mean values and high variances in the random variables. This may be viewed from Figure 3.2.

Figure 3.2: Normal vs Lognormal Distribution: Graphical Shapes



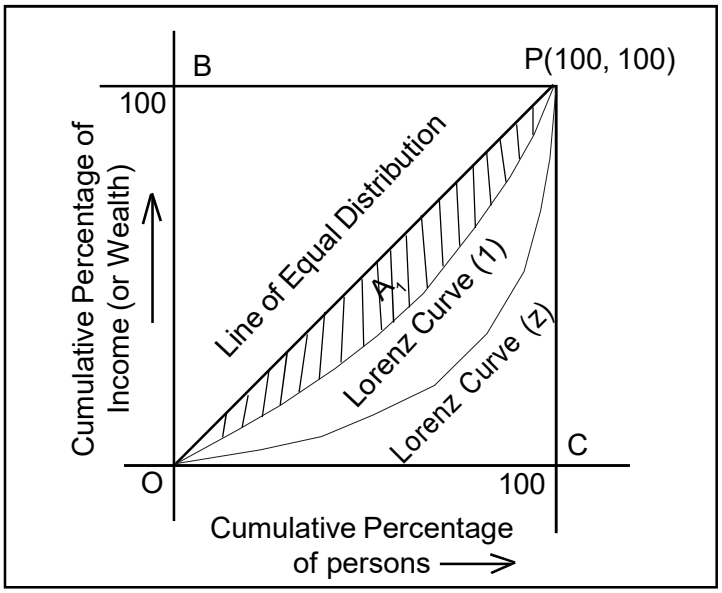
3.5 LORENZ CURVE

The graphical method of measurement of dispersion was developed by a famous economic statisticion Dr. Max O. Lorenz. He used this method to measure the inequalities of income or wealth of a society. In this method, cumulative percentage of income or wealth, i.e., the percentage of income (or wealth) less than a given value is plotted against the cumulative percentage of persons. The curve thus obtained, is known as Lorenz Curve. Another name of Lorenz Curve is 'Cumulative percentage curve'.

In figure, the cumulative percentage of persons is shown on X-axis and cumulative percentage of other characteristics (income or wealth) is shown on Y-axis)

In Figure 3.3 a point P with coordinates (100, 100) is plotted on the figure and perpendiculars PB and PC are dropped on the Y-axis and X-axis respectively from this point. The line joining the point P and O represents the line of equal distribution. Any departure from this line denotes the extent of inequality. For example, the extent of inequalities, in the above figure, are represented by the area A_1 between the Lorenz Curve and the line of equal distribution OP.

Figure 3.3: Lorenz Curve



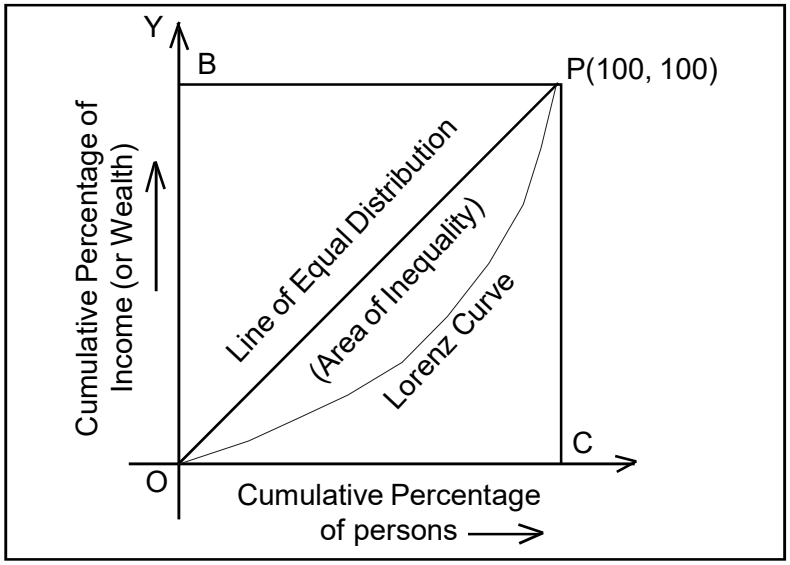
3.6 COEFFICIENT OF INEQUALITY : GINI COEFFICIENT

Gini coefficient is based on the Lorenz Curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half-square in which the curve lies.

Let the area of the triangle POC be denoted by A. The coefficient of inequality i.e. the Gini coefficient is defined by the ratio $\frac{A_1}{A}$, where A_1 is the area between the Lorenz curve (1) and the line of equal distribution OP. Higher the degree of inequality higher is the ratio and accordingly hogher is the value of the Gini coefficient.

The value of the Gini coefficient can theoretically be between the two extreme values of O (indicating perfect equality) to 1 (indicating perfect inequality). This has been shown in Figure 3.4.

Figure 3.4: Lorenz Curve and Gini Coefficient



Example : Use Lorenz Curve to compare the extent of inequalities of income distribution in the two groups.

Monthly Income (Rs.)	No. of Persons in	
	Group A	Group B
1200–1400	800	4800
1400–1600	960	6400
1600–1800	1040	9600
1800–2000	600	3600
2000–2200	480	8000
2200–2400	120	4000

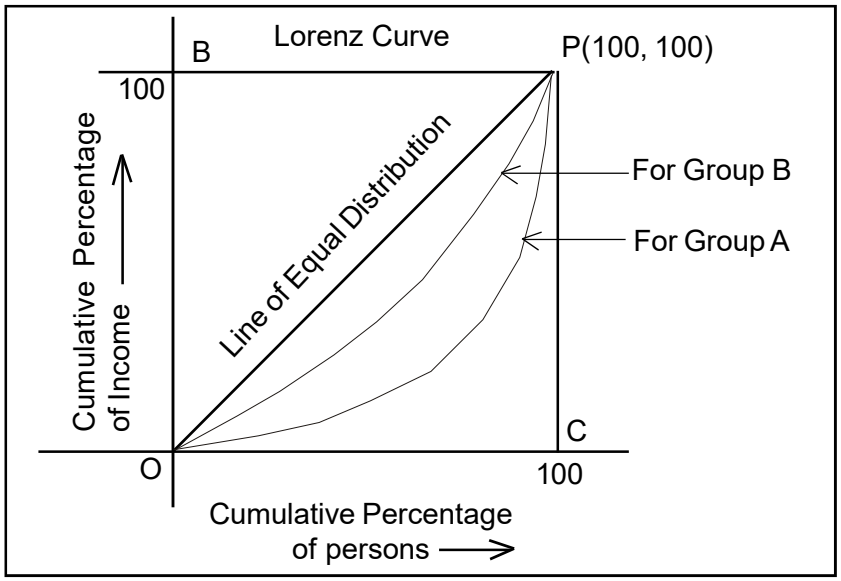
Solution : Let us prepare the following table :

Monthly Income (Rs.)	Mid Values (Rs.)	Cum. Income	Cum. %	Group A			Group B		
				No. of persons	Cum. Total	Cum. %	No of persons	Cum. Total	Cum. %
1200–1400	1300	1300	12	800	800	20	4800	4800	13
1400–1600	1500	2800	26	960	1760	44	6400	11200	31
1600–1800	1700	4500	42	1040	2800	70	9600	20800	57
1800–2000	1900	6400	59	600	3400	85	3600	24400	67
2000–2200	2100	8500	79	480	3880	97	8000	32400	89
2200–2400	2300	10800	100	120	4000	100	4000	36400	100

Note : The percentages are approximated to the nearest whole number.

This may be represented with the help of Figure 3.5.

Figure 3.5: Income inequalities between Group A & B



From Figure 3.5 it is obvious from the above figure that inequalities in the distribution of income are more in group A than in group B.



CHECK YOUR PROGRESS

Q 1: How did Pareto formulated the law of income distribution? (Answer in about 60 words).

.....

.....

.....

.....

.....

Q 2: What does the Lorenz curve reflect upon? (Answer in about 40 words).

.....

.....

.....

.....

Q 3: Define the Gini Coefficient. (Answer in about 30 words).

.....

.....

.....



3.7 LET US SUM UP

- The concept of Pareto Distribution was put forward by Vilfredo Pareto. This is also referred to as the Pareto Principle or the 80-20 Rule. The Pareto distribution is used in describing social, scientific, and geophysical phenomena in a society. Pareto created a mathematical formula in the early 20th century that described the inequalities in wealth distribution that existed in his native country of Italy.
- The definition of the Pareto Distribution was later expanded in the 1940s by Dr. Joseph M. Juran, a prominent product quality guru. Juran applied the Pareto principle to quality control for business production to show that 20 percent of the production process defects are responsible for 80 percent of the problems in most products.
- A log-normal distribution is a statistical distribution of logarithmic values from a related normal distribution. A log-normal distribution can be translated to a normal distribution and vice versa using associated logarithmic calculations.
- The graphical method of measurement of dispersion was developed by a famous economic statisticion Dr. Max O. Lorenz. He used this method to measure the inequalities of income or wealth of a society. In this method, cumulative percentage of income or wealth, i.e., the percentage of income (or wealth) less than a given value is plotted against the cumulative percentage of persons. The curve thus obtained, is known as Lorenz Curve.
- Gini coefficient is based on the Lorenz Curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half-square in which the curve lies.



3.8 FURTHER READING

1. A.M. Gun, M.K. Gupta, B. Dasgupta, "Fundamentals Of Statistics", Volume two, Seventh Edition.
2. Richard I. Levin, David S. Rubin,"Statistics For Management", Seventh Edition.



3.9 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: In 1906, Vilfredo Pareto observed that 20 percent of the pea pods were responsible for 80 percent of the peas planted in his garden. Again, he also found that 80 percent of his country's wealth was owned by about 20 percent of its population. Further, he also found that 80 percent of the land was owned by a handful of wealthy citizens, who comprised about 20 percent of the population. On the basis of such observations, Paretor formulated his law of income distribution.

Ans to Q No 2: Lorenz curves tries to measure the inequalities of income or wealth of a society. In this method, cumulative percentage of income or wealth, i.e., the percentage of income (or wealth) less than a given value is plotted against the cumulative percentage of persons.

Ans to Q No 3: Gini coefficient is based on the Lorenz Curve and is defined as the ratio of the area between the diagonal and the Lorenz curve to the total area of the half-square in which the curve lies.



3.10 MODEL QUESTIONS

Answer the following question in around 150 words

Q 1: Describe Pareto's law of income distribution and its empirical validity.

Q 2: Describe the concept of Lorenz curve.

Q 3: Discuss the concept of log-normal distribution. Mention its vaious uses.

*** ***** ***

UNIT 4 : CORRELATION

UNIT STRUCTURE

4.1 Learning Objectives

4.2 Introduction

4.3 Correlation Analysis

4.4 Types of Correlation

4.5 Methods of Measuring Correlation

4.6 Properties of Correlation Coefficient

4.7 Rank Correlation Coefficient

4.8 Tied or Repeated Ranks

4.9 Multiple Choice Questions

4.10 Exercises

4.11 Let Us Sum Up

4.12 Further Reading

4.13 Answers to Check Your Progress

4.14 Model Questions

4.1 LEARNING OBJECTIVES

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

- put forward the meanings of bivariate data
- state the concept of correlation between two variables and quantitative measurement of correlation including the interpretation of positive, negative and zero correlation
- use the technique of finding correlation between two quatitative characteristics and its interpretation.

4.2 INTRODUCTION

In the previous chapters, we discussed many statistical measures relating to univariate distribution, i.e. distribution of one variable like height, weight, mark, profit, wages and so on. However, there are situations that demand study of more than one variable simultaneously. A distribution of

paired values of two variables is called a bivariate distribution and a distribution of the values of more than two variables is called a multivariate distribution. If x and y denote marks in Mathematics and Statistics for a group of 30 students, then the corresponding bivariate data would be (x_i, y_i) for i = 1, 2, . . ., 30 where (x_i, y_i) denotes the marks in Mathematics and Statistics for the student with serial number i and so on.

If two variables are such that when one changes the other also changes then the two variables are said to be correlated. In the practical field we have to study the type of relationship that might exist between the ages of husbands and wives, the heights of fathers and sons, the volume of production of a particular crop, the price of a commodity and the demand for it, and so on. When we study the relationship between two variables only then it is called simple correlation or correlation.

4.3 CORRELATION ANALYSIS

Which studying two variables at the same time, if it is found that the change in one variable is reciprocated by a corresponding change in the other variable either directly or inversely, then the two variables are known to be associated or correlated. Otherwise, the two variables are known to be dissociated or uncorrelated or independent. For example, if the price of a commodity falls, the demand for it rises; if price rises, demand of that commodity falls. Thus there is correlation between the two variables price and demand.

Causation and Correlation

Causation or causality, is the capacity of one variable to influence another. The first variable may bring the second into existence or may cause the incidence of the second variable to fluctuate.

Causation is often confused with correlation, which indicates the extent to which two variables tend to increase or decrease in parallel. However, correlation by itself does not imply causation. There may be a third factor, for example, that is responsible for the fluctuations in both variables.

A statistically significant correlation has been reported, for example, between yellow cars and a lower incidence of accidents. That does not

indicate that yellow cars are safer, but just that fewer yellow cars are involved in accidents. A third factor, such as the personality type of the purchaser, is more likely to be responsible.

4.4 TYPES OF CORRELATION

Correlation for two variables may be classified into the following three types :

(i) Positive correlation,

(ii) Negative correlation and

(iii) Zero correlation

(i) Positive correlation : If two related variables move in the same derection i.e. an increase (or decrease) on the part of one variable results an increase (or decrease) on the part of the other variable, then the two variables are said to be positively correlated. As for example, height and weight yield, profit and investment etc. are positively correlated.

When the changes in two related variables are exactly proportional and are in the same direction then we say that there is perfect positive correlation between them. For example, when temparature of a gas increases, the volume of the gas also increases at the same rate. In this case temparature and volume are perfectly positively correlated.

(ii) Negative correlation : If two related variables move in the opposite directions i.e. an increase (or a decrease) on the part of one variable results a decrease (or an increase) on the part of the other variable, then the two variable are known to have a negative carrelation. For example if price of a certain commodity increases(decreases) then in normal situation, demand for that commodity decreases(increases). Therefore, price and demand of a commodity is negatively correlated.

(iii) Zero correlation : Two variables are said to have zero correlation or no correlation between them if they tend to change with no connection to each other. In this case the variables are said to be uncorrelated. As for example, shoe size and intelligence are said to have zero correlation.

4.5 METHODS OF MEASURING CORRELATION

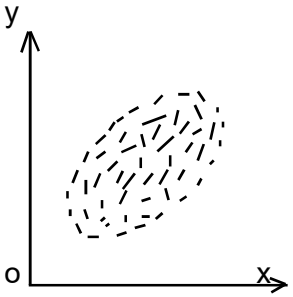
There are basically two methods of measuring correlation between two variables, viz

(i) Diagramatic method and

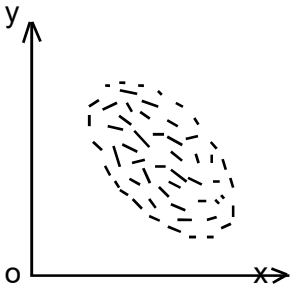
(ii) Mathematical method

(i) Diagramatic method : The diagramatic method is also known as the scatter diagram method. In this method, we can have an idea about the extent of association between two variables graphically.

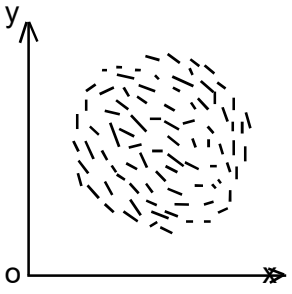
Yet the bivariate data be denoted by (x_i, y_i), where i = 1, 2, . . ., n. If these paired values are plotted on a graph paper by taking the values of x on the x -axis and the values of y on the y -axis, then the diagram obtained on the graph paper is called a scatter diagram.



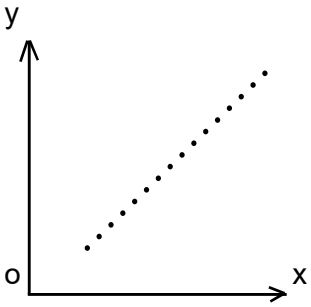
(a) Positive correlation



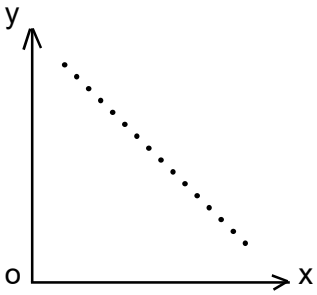
(b) Negative correlation



(c) Zero correlation



(d) Perfect positive correlation



(e) Perfect Negative correlation

If the plotted points lie from lower left corner to upper right corner (as shown in fig. (a)) then it indicates that there is positive correlation between the two variables.

If the plotted point concentrate from upper left to lower right (as shown in fig. (b)) then it reveals that there is negative correlation between the two variables.

If the plotted points fall haphazardly without depicting any particular pattern (as shown in fig. (c)) then it indicates that there is no correlation or zero correlation between the two variables.

If all the points in the scatter diagram fall on a straight line with positive slope (as shown in fig. (d)) then it implies that there is perfect positive correlation between the two variables.

Again, if all the points fall on a straight line with negative slope (as shown in fig. (e)) then it indicates that there is perfect negative correlation between the two variables.

(ii) Mathematical Method : The following are mathematical methods of measuring correlation between two variables :

(a) Karl Pearson's coefficient of correlation

(b) Spearman's rank correlation coefficient

(a) Karl Pearson's Coefficient of Correlation : Karl Pearson's coefficient of correlation is a statistical tool of measuring degree linear relationship between two variables numerically. The coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviations of the two variables. If the two variables are denoted by x and y and if the corresponding bivariate data are (x_i, y_i), for i = 1, 2, 3, . . ., n, then the coefficient of correlation between x and y, due to Karl Pearson, is given by :

r or $r_{xy} = \frac{Cov(x, y)}{\sigma_x . \sigma_y}$

Where $Cov(x, y) = \frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y})$

$= \frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}$ (after simplification)

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

$= \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}$ (after simplification)

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

$$= \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2} \qquad \text{(after simplification)}$$

Thus, the simplified form of the coefficient of correlation is given as:

$$r_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}}$$

$$= \frac{\frac{1}{n}\Sigma xy - \frac{\Sigma x}{n}.\frac{\Sigma y}{n}}{\sqrt{\frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2}\sqrt{\frac{1}{n}\Sigma y^2 - \left(\frac{\Sigma y}{n}\right)^2}}$$

$$= \frac{\frac{1}{n}\left[\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right]}{\sqrt{\frac{1}{n}\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right].\frac{1}{n}\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

$$= \frac{\frac{1}{n}\left[\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right]}{\sqrt{\left(\frac{1}{n}\right)^2\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

$$= \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

Assumed Mean Method

When deviations are taken from assumed means, i.e. if we assume $d_x = x - A, d_y = y - B$, A and B are assumed means of x and y respectively, then the following formula is used.

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\sqrt{\left[\Sigma {d_x}^2 - \frac{(\Sigma d_x)^2}{n}\right]\left[\Sigma {d_y}^2 - \frac{(\Sigma d_y)^2}{n}\right]}}$$

4.6 PROPERTIES OF CORRELATION COEFFICIENT

1. The coefficient of correlation is a unit-free measure.

   This means that if x denotes height of a group of students expressed in cm and y denotes their weight expressed in kg, then the correlation coefficient between height and weight would be free from any unit.

2. The coefficient of correlation remains invariant under a change of origin and / or scale of the variables under consideration.

   This property states that if the original pair of variables x and y is changed to a new pair of variables u and v by effecting a change of origin and scale for both x and y i.e.

   $u=\frac{x-a}{b}$ and $v=\frac{y-c}{d}$

   Where a and c are the origins of x and y and b and d are the respective scales, then we have $r_{xy}=\frac{bd}{/b//d/}r_{uv}$

3. The coefficient of correlation always lies between -1 and 1, including both the limiting values i.e. $-1\le r\le 1$.

Example 1 : Compute the correlation coefficient between x and y from the following data :

$$n=10,\ \Sigma xy=220,\ \Sigma x^2-200,\ \Sigma y^2-262,\ \Sigma x=40,\ \Sigma y=50$$

Solution : We know that,

$$r_{xy}=\frac{\Sigma xy-\frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2-\frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2-\frac{(\Sigma y)^2}{n}\right]}}$$

$$=\frac{220-\frac{40\times 50}{10}}{\sqrt{\left[200-\frac{40^2}{10}\right]\left[262-\frac{50^2}{10}\right]}}$$

$$=\frac{220-200}{\sqrt{(200-160)(262-250)}}$$

$$= \frac{20}{\sqrt{40 \times 12}}$$

= 0.91 (approx)

Example 2: Find Karl Pearsons correlation coefficient from the following information :

X :	2	3	5	5	6	8
Y :	9	8	8	6	5	3

Solution : Let us prepare the following table :

x	y	xy	x^2	y^2
2	9	18	4	81
3	8	24	9	64
5	8	40	25	64
5	6	30	25	36
6	5	30	36	25
8	3	24	64	9
$\Sigma x = 29$	$\Sigma y = 39$	$\Sigma xy = 166$	$\Sigma x^2 = 163$	$\Sigma y^2 = 279$

Now,

$$r_{xy} = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]}\sqrt{\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

$$= \frac{166 - \frac{29 \times 39}{6}}{\sqrt{\left[163 - \frac{29^2}{6}\right]}\sqrt{\left[279 - \frac{39^2}{6}\right]}}$$

= 0.93 (approx) (after simplification)

Example 3 : Find out the coefficient of correlation between the per capita income and the price level from the following data :

Year :	1980	1981	1982	1983	1984	1985
Per capita income :	360	420	500	556	600	590
Price index :	100	104	115	160	280	290

Solution : Let, x denotes per capita income and y denotes the price level or price index.

Now, we prepare the following table :

x	y	d_x = x-A = x-500	d_y = y-B = y-160	dx^2	dy^2	dx.dy
360	100	–140	–60	19600	3600	8400
420	104	–80	–56	6400	3136	4480
500	115	0	–45	0	2025	0
556	160	56	0	3136	0	0
600	280	100	120	10000	14400	12000
590	290	90	130	8100	16900	11700
		$\Sigma dx = 26$	$\Sigma dy = 89$	$\Sigma dx^2 = 47236$		$\Sigma d_x d_y = 36580$

Now using formula for assumed mean method we have,

$$r_{xy} = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\sqrt{\left\{\Sigma d^2{}_x - \frac{(\Sigma d_x)^2}{n}\right\}}\sqrt{\left\{\Sigma d^2{}_y - \frac{(\Sigma d_y)^2}{n}\right\}}}$$

$$= \frac{36580 - \frac{26 \times 89}{6}}{\sqrt{\left\{47236 - \frac{26^2}{6}\right\}}\sqrt{\left\{40061 - \frac{89^2}{6}\right\}}}$$

$$= \frac{36580 - 385.67}{\sqrt{47123.33 \times 38740.83}}$$

$$= \frac{36194.33}{42727.87}$$

= 0.847 (approx)

Example 4 : Determine the value of the correlation coefficient r from the following data :

$$\sum_{i=1}^{8}(x_i - \bar{x})^2 = 32, \qquad \sum_{i=1}^{8}(y_i - \bar{y})^2 = 72$$

$$\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y}) = 43$$

Solution : Now, $Cov(x,y) = \frac{1}{n}\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{8}\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{8} \times 43 \quad = 5.375$$

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

$$= \sqrt{\frac{1}{8}\sum_{i=1}^{8}(x_i - \bar{x})^2}$$

$$= \sqrt{\frac{1}{8} \times 32}$$

$$= \sqrt{4} \qquad = 2$$

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

$$= \sqrt{\frac{1}{8}\sum_{i=1}^{8}(y_i - \bar{y})^2}$$

$$= \sqrt{\frac{1}{8} \times 72}$$

$$= \sqrt{9} \qquad = 3$$

$$\therefore r = \frac{Cov(x,y)}{\sigma x . \sigma y}$$

$$= \frac{5.375}{2 \times 3}$$

$$= \frac{5.375}{6}$$

= 0.8958

Exmple 5: Calculate the Karl Pearson's Coefficient of correlation from the following data :

(i) Sum of deviations of x values = 5

(ii) Sum of deviations of y values = 4

(iii) Sum of squares of deviations of x values = 40

(iv) Sum of squares of deviations of y values = 50

(v) Sum of the product of deviations of x and y values = 32

(vi) No. of pairs of observations = 10

Solution : Let, d_x=X–A and d_y=Y–b be the deviations of X and Y values.

We are given, $\sum d_x = 5,\ \sum d_y = 4$

$$\sum {d_x}^2 = 40,\ \sum {d_y}^2 = 50$$

$$\sum d_x \cdot d_y = 32 \text{ and } n = 10$$

Now, using formula for assumed mean method, we have,

$$r_{xy} = \frac{\sum d_x d_y - \frac{\left(\sum d_x\right)\left(\sum d_y\right)}{n}}{\sqrt{\left\{\sum {d_x}^2 - \frac{\left(\sum d_x\right)^2}{n}\right\}\left\{\sum {d_y}^2 - \frac{\left(\sum d_y\right)^2}{n}\right\}}}$$

$$= \frac{32 - \frac{5 \times 4}{10}}{\sqrt{\left\{40 - \frac{5^2}{10}\right\}\left\{50 - \frac{4^2}{10}\right\}}}$$

$= 0.704$ (after simplification)

Example 6: From the following data compute coefficient of correlation between X and Y.

	X Series	Y Series
Arithmetic Mean (AM)	25	18
Sum of Squares of deviations from AM	136	138

Summation of products of deviations of X and Y series from their respective means=122

Number of pairs of values = 15

Solution : We are given,

$$\bar{x} = 25,\ \bar{y} = 18$$

$$\sum \left(x - \bar{x}\right)^2 = 136,\ \sum \left(y - \bar{y}\right)^2 = 138$$

$$\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right) = 122,\ n = 15$$

$$\therefore r_{xy} = \frac{Cov(x, y)}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\sqrt{\sum \left(x - \bar{x}\right)^2 \sum \left(y - \bar{y}\right)^2}}$$

$$= \frac{122}{\sqrt{136 \times 138}} = \frac{122}{137} = 0.89$$



CHECK YOUR PROGRESS

Q.1: Fill in the blanks :

(i) Correlation coefficient is _____ of the units of measurement.

(ii) Coefficient of correlation is a measure of the strength of the _____relationship between two variables.

(iii) Coefficient of correlation lies between _____ and _____

(iv) If two variables are independent they are_____.

4.7 RANK CORRELATION COEFFICIENT

Rank correlation coefficient is a method of usually measuring the relationship between two qualitative variable or attributes such as honesty, beauty, morality, etc. which cannot be measured numerically. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks. Charles Edward Spearman, a British psychologist developed the rank correlation coefficient formula and hence rank correlation coefficient is also known as spearman's rank correlation coefficient.

Spearman's rank correlation coefficient is denoted by R or Y and is given as $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.

Note : The value of R always lees between – 1 and 1, i.e. $-1 \le R \le 1$.

Example 7 : Compute rank correlation from the following data relating to ranks given by two judges in a contest.

Serial No. of candidate :	1	2	3	4	5	6	7	8	9	10
Rank by Judge A :	10	5	6	1	2	3	4	7	8	9
Rank by Judge B :	5	6	9	2	8	7	3	4	10	1

Solution : Let R_1 denotes the ranks given by Judge A and R_2 denotes the ranks given by Judge B.

Now, let us prepare the following table :

Serial No.	R_1	R_2	$d = R_1 - R_2$	d^2
1	10	5	5	25
2	5	6	–1	1
3	6	9	–3	9
4	1	2	–1	1
5	2	8	–6	36
6	3	7	–4	16
7	4	3	1	1
8	7	4	3	9
9	8	10	–2	4
10	9	1	8	64
				$\Sigma d^2 = 166$

$\therefore$ The rank correlation coefficient is given by,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 166}{10(10^2-1)}$$

= 0.006

Example 8 : Compute the coefficient of rank correlation between sales and advertisement expressed in thousands of rupees from the following data :

Sales :	90	85	68	75	82	80	95	70
Advertisement :	7	6	2	3	4	5	8	1

Solution : Let R_1 denotes ranks for sales and R_2 denotes ranks for advertisement.

Now, let us prepare the following table :

Sales	Advertisement	Ranks for Sales(R_1)	Ranks for Advertisement (R_2)	d = R_1–R_2	d^2
90	7	2	2	0	0
85	6	3	3	0	0
68	2	8	7	1	1
75	3	6	6	0	0
82	4	4	5	–1	1
80	5	5	4	1	1
95	8	1	1	0	0
70	1	7	8	–1	1
					$\Sigma d^2 = 4$

∴ Rank correlation coefficient,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 4}{8(8^2-1)}$$

= 1 – 0.0476

= 0.95

Example 9: The following are the marks obtained by a group of students in two papers. Calculate the rank coefficient of correlation.

Economics :	78	36	98	25	75	82	92
Statistics :	84	51	91	69	68	62	86
Economics :	62	65	39				
Statistics :	58	35	49				

Solution : Let R_1 denotes the ranks of marks obtained in Economics and R_2 denotes the ranks of marks obtained in Statistics.

Now, let us prepare the following table :

Marks in Economics	Marks in Statistics	R_1	R_2	d=R_1–R_2	d^2
78	84	4	3	1	1
36	51	9	8	1	1
98	91	1	1	0	0

25	69	10	4	6	36
75	68	5	5	0	0
82	62	3	6	–3	9
92	86	2	2	0	0
62	58	7	7	0	0
65	35	6	10	–4	16
39	49	8	9	–1	1
					$\Sigma d^2 = 64$

$\therefore$ Rank correlation coefficient,

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 64}{10(10^2-1)}$$

= 1 – 0.388

= 0.612 (approx)

4.8 TIED OR REPEATED RANKS

In case of a tie, i.e., when two or more individuals have the same rank, each individual is assigned a rank equal to the mean of the ranks.

In the formula for tied rank we are to add the factor $\frac{m(m^2-1)}{12}$ to Σd^2 where m is the number of times an item is repeated. The correction factor is to be added for each repeated value in both the series. Thus spearmain's rank correlation coefficient formula for tied ranks becomes,

$$R = 1 - \frac{6\left[\Sigma d^2 + \Sigma \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

Example 10: Calculate the spearman's ranks correlation coefficient between the series A and B given below :

Series A :	57	59	62	63	64	65	55	58	57
Series B :	113	117	126	126	130	129	111	116	112

Solution : Let, R_1 denotes ranks of the series A and

R_2 denotes ranks of the series B

Now, let us prepare the following table :

Series A	Series B	Rank of Series A(R_1)	Rank of Series B(R_2)	d = R_1-R_2	d^2
57	113	2.5	3	–0.5	0.25
59	117	5	5	0	0
62	126	6	6.5	–0.5	0.25
63	126	7	6.5	0.5	0.25
64	130	8	9	–1	1.00
65	129	9	8	1	1.00
55	111	1	1	0	0
58	116	4	4	0	0
57	112	2.5	2	0.5	0.25
					$\Sigma d^2 = 3.00$

The observation 57 in series A is repeated twice and the observation 126 in series B is also repeated twice.

$$\therefore \Sigma\frac{m(m^2-1)}{12} = \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} \quad \text{(where } m_1 = 2,\ m_2 = 2\text{)}$$

$$= \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12}$$

$$= \frac{4-1}{6} + \frac{4-1}{6}$$

$$= \frac{3}{6} + \frac{3}{6}$$

$$= \frac{6}{6} \quad = 1$$

$\therefore$ Rank correlation coefficient

$$R = 1 - \frac{6\left[\Sigma d^2 + \Sigma\frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\,[3+1]}{9(9^2-1)}$$

$$= 1 - \frac{6 \times 4}{9 \times 80}$$

= 0.967



CHECK YOUR PROGRESS

Q.2: Fill in the blanks :

(i) Rank correlation coefficient lies between _____/

(ii) In rank correlation coefficient the association need not be_____.

(iii) The sum of the difference of rank is_____.

(iv) Great advantage of_____ is that it can be used to rank attributes which can not be expressed by way of numerical value.

4.9 MULTIPLE CHOICE QUESTIONS

Q 1: If r_{xy} = 0, the variables x and y are

(a) linearly related (b) independent

(c) not linearly related (d) none of the above

Ans : (c)

Q 2: If $r_{xy} = \pm 1$, the two lines of regression are

(a) coincident (b) paralled

(c) perpendicular to each other (d) none of the above

Ans : (a)

Q 3: For finding correlation between two attributes, we consider

(a) Pearson's correlation coefficient

(b) Scatter diagram

(c) Spearman's rank correlation coefficient

(d) All these

Ans : (c)

Q 4: If the relationship between two variables x and y is given by $2x + 3y + 4 = 0$, then the value of the correlation coefficient between x and y is

(a) 0 (b) 1

(c) –1 (d) negative

Ans : (c)

Q 5: The coefficient of correlation between two variables

(a) can have any unit

(b) is expressed as the product of units of the two variables

(c) is a unitfree measure

(d) none of these

Ans : (c)

Q 6: Scatter diagram is considered for measuring

(a) Linear relationship between two variables

(b) Curvitinear relationship between two variables

(c) Neither (a) nor (b)

(d) Both (a) and (b)

Ans : (d)

Q 7: The correlation between the speed of an automobile and the distance travelled by it after applying the brakes is

(a) Negative (b) Zero

(c) Positive (d) None of these

Ans : (a)

Q 8: Pearson's correlation coefficient is used for finding

(a) Correlation for any type of relation

(b) Correlation for linear relation only

(c) Correlation for curvilinear relation only

(d) Both (b) and (c)

Ans : (b)

Q 9: If the plotted points in a scatter diagram are evenly distributed, then the correlation is

(a) Zero (b) Negative

(c) Positive (d) (a) or (b)

Ans : (a)

Q 10: In rank correlation coefficient the association need not be linear

(a) true (b) false

(c) both (d) none

Ans : (a)

4.10 EXERCISES

Q 1: What is the value of correlation coefficient due to Pearson on the basis of the following data :

x :	–5	–4	–3	–2	–1	0	1	2	3	4	5
y :	27	18	11	6	3	2	3	6	11	18	27

Ans : Zero

Q 2: If for two variables x and y, the covariance, variance of x and variance of y are 40, 16 and 256 respectively, what is the value of the correlation coefficient ?

Ans : 0.625

Q 3: What is the coefficient of correlation from the following data ?

x :	1	2	3	4	5
y :	8	6	7	5	5

Ans : –0.85

Q 4: Given that for twenty pairs of observations,

$\Sigma xy = 525,\ \Sigma x = 129,\ \Sigma y = 97,\ \Sigma x^2 = 687$

$\Sigma y^2 = 427$ and $u = 10 - 3y$. Find the coefficient of correlation between x and u .

Ans : –0.74

Q 5: Eight contestants in a musical contist were ranked by two judges A and B in the following manner :

Serial number of the contestants :	1	2	3	4	5	6	7	8
Rank by Judge A :	7	6	2	4	5	3	1	8
Rank by Judge B :	5	4	6	3	8	2	1	7

Find the rank correlation coefficient.

Ans : 0.57

Q 6: Find the value of Rank correlation coefficient between the following marks in Economics and Statistics.

Roll No. :	1	2	3	4	5	6
Marks in Economics :	25	30	46	30	55	80
Marks in Statistics :	30	25	50	40	50	78

Ans : 0.857

Q 7: Discuss the different types of correlation.

Q 8: Write short notes on :

(a) Rank correlation

(b) Scatter diagram

(c) Positive and Negative correlation.

Q 9: From the followings data, calculate Karl Pearson's coefficient of correlation :

(i) $\sum_{i=1}^{9}(x_i-\bar{x})^2=120$, $\sum_{i=1}^{9}(y_i-\bar{y})^2=346$, $\sum_{i=1}^{9}(x_i-\bar{x})(y_i-\bar{y})=193$

(ii) $\Sigma x=125,\ \Sigma y=80,\ n=10,\ \Sigma x^2=1586,\ \Sigma y^2=650,\ \Sigma xy=1007$

(iii) $n=10,\ \Sigma x=140,\ \Sigma y=150,\ \Sigma(x-14)^2=180$,

$\Sigma(y-15)^2=215,\ \Sigma(x-14)(y-15)=60$

Ans : (i) 0.95, (ii) 0.47, (iii) 0.92



4.11 LET US SUM UP

- If two variables are such that when one changes the other also changes then the two variables are said to be correlated.
- Correlation for two variables may be classified into the following three types (i) Positive correlation,
  (ii) Negative correlation and
  (iii) Zero correlation
- If two related variables move in the same derection i.e. an increase (or decrease) on the part of one variable results an increase (or decrease) on the part of the other variable, then the two variables are said to be positively correlated.

- If two related variables move in the opposite directions i.e. an increase (or a decrease) on the part of one variable results a decrease (or an increase) on the part of the other variable, then the two variable are known to have a negative carrelation.
- Two variables are said to have zero correlation or no correlation between them if they tend to change with no connection to each other.
- There are basically two methods of measuring correlation between two variables, viz (i) Diagramatic method and
  (ii) Mathematical method
- The following are mathematical methods of measuring correlation between two variables :
  (a) Karl Pearson's coefficient of correlation
  (b) Spearman's rank correlation coefficient
- Karl Pearson's coefficient of correlation is a statistical tool of measuring degree linear relationship between two variables numerically. The coefficient may be defined as the ratio of covariance between the two variables to the product of the standard deviations of the two variables.
- Rank correlation coefficient is a method of usually measuring the relationship between two qualitative variable or attributes such as honesty, beauty, morality, etc. which cannot be measured numerically.
- Spearman's rank correlation coefficient is denoted by R or Y and is given as $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

  Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.
- In case of a tie, i.e., when two or more individuals have the same rank, each individual is assigned a rank equal to the mean of the ranks.



4.12 FURTHER READING

1) Gujrati, D (1995); Basic Econometrics (third edition), McGraw-Hill, Inc; Singapore

2) Gun, A M; Gupta, M K & Dasgupta, B (2002); Fundamentals of Statistics, Volume One; World Press Private Ltd. Kolkata

3) Gupta, S C (1994); Fundamentals of Statistics; Himalayan Publishing House, New Delhi.

4.13 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q. No. 1 : (i) independent (ii) linear
(iii) -1, 1 (iv) uncorrelated

Ans to Q. No. 2 : (i) -1 to +1 (ii) linear
(iii) zero (iv) rank correlation

4.14 MODEL QUESTIONS

A. SHORT TYPE QUESTIONS

Q 1: Give the concept of correlation.

Q 2: Distinguish between causation and correlation.

Q 3: What is meant by zero correlation?

Q 4: Write down the properties of correlation coefficcient.

B. LONG QUESTIONS

Q 1: Find the coefficient of correlation from the given data in each of the following cases.

(i) x :	6	8	15	18	20	24	28	31		
y :	10	12	15	18	25	22	26	28		
(ii) x :	39	65	62	90	82	75	25	98	36	78
y :	47	53	58	86	62	68	60	91	51	84
(iii) x :	9	8	7	6	5	4	3	2	1	
y :	15	16	14	12	10	13	11	8	9	

Ans : (i) 0.96, (ii) 0.78, (iii) 0.88

Q 2: Calculate rank correlation coefficient from the following data :

x :	75	88	95	70	60	80	81	50
y :	120	134	150	115	110	140	142	100

Ans : 0.92

Q 3: Discuss various methods of measuring correlation.

Q 4: The following table provides the distribution of items according to size groups and also the number of defectives :

Size group :	9-11	11-13	13-15	15-17	17-19
No. of items :	250	350	400	300	150
No. of defective items :	25	70	60	45	20

Find the correlation coefficient between size and defectives.

Ans : 0.07

*** ***** ***

UNIT 5 : REGRESSION ANALYSIS

UNIT STRUCTURE

5.1 Learning Objectives

5.2 Introduction

5.3 Regression

5.4 Two-variable Regression Model (Least Square Method)

5.5 Properties of Regression Coefficients

5.6 Correlation vs. Regression

5.7 Multiple Choice Questions

5.8 Exercises

5.9 Let Us Sum Up

5.10 Further Reading

5.11 Answers to Check Your Progress

5.12 Model Questions

5.1 LEARNING OBJECTIVES

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

- put foward the concept of regression, regression equations and regression lines
- state the different kinds of regression and their applications
- utilise the technique of estimation of a variable from known set of data.

5.2 INTRODUCTION

If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1, 2, . . ., n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation in Statistics.

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. The nature

of a regression equation is different from the nature of a mathematical equation, e.g., if Y = 10 + 2x is a mathematical equation then it implies that Y is exactly equal to 20 when X = 5. Howerver, if Y = 10 + 2x is a regression equation, then Y = 20 is an average or most likely or most probable value of Y when X = 5. The variable corresponding to whose value the average value of the other variable is estimated is called the independent or explanatory or exogenous variable. The variable whose average value is estimated is called the dependent or exphained or endogenous variable.

5.3 REGRESSION

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data. In other words, regression is a technique of estimating the most probable or most likely or average values of one variable for specified values of the other variable on the basis of the given values of the two variables. This is done by estimating a relationship between the two variables.

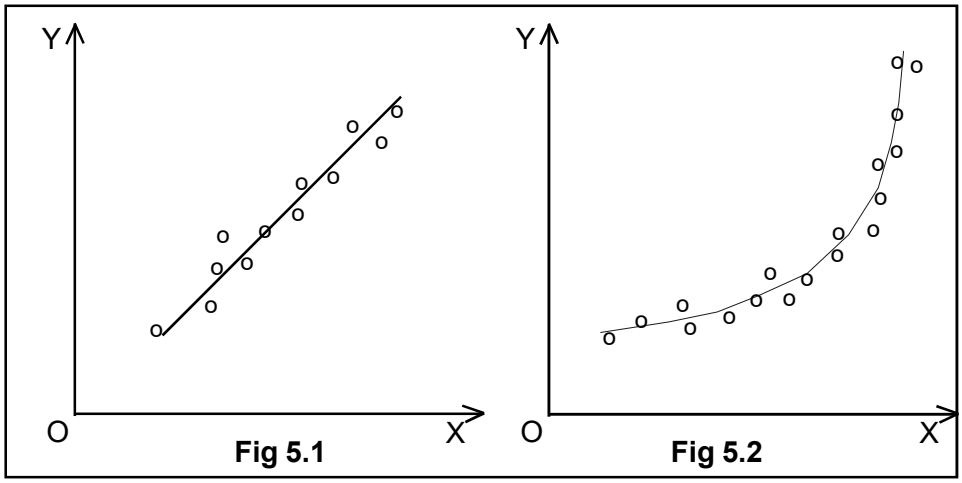
Regression Equations

The two equations which are estimated on the basis of the given values of the two variables to estimate the most likely values of one variable for specified values of the other are called regression equations. The geometrical representations of regression equations are called regression lines. The terms regression equations and regression lines are used interchangeably.

Linear and nonlinear or curvilinear regression

If the points of the scatter diagram of two correlated variables clusters around a straight line then we say that there is linear regreassion between the two variables. If the scatter diagram clusters around a curve then we say that these is nonlinear or curvilinear regression between the two variables. For example let x denotes the height of n persons and y denotes their weight. Yet the values of x be $x_1, x_2, ..., x_n$ and the corresponding values of y be $y_1, y_2, ..., y_n$. Now from these bivariate data we get the ordered pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, each of which represents a point in the xy - plane. If these point cluster around a line as shown in figure 5.1 then it

indicates that there is a linear regression between height (x) and weight (y) and if the point cluster around a curve as shown in figure 5.2 then it indicates that there is non-linear or curvilinear regression between x and y .



5.4 TWO-VARIABLE REGRESSION MODEL (LEAST SQUARE MEHTOD)

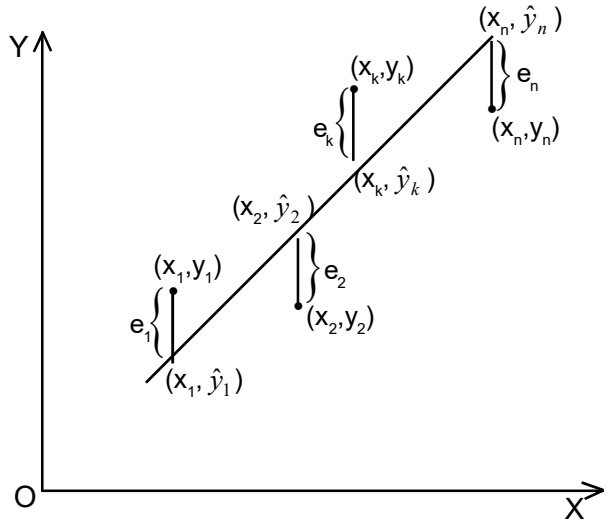
When only two variables are associated with regression analysis then we call it single regression. When we say regression it means simple regression. In case of a simple regression model if depends on x , then the regression model is termed as regression lines of y on x and is given by,

Y=a+bx -(1)

Here a and b are two constants and they are also known as regression parameters. Furthermore b is also known as the regression coefficient of y on x and is also denoted by b_{yx}. We may define the regression line of y on x as the line of best fit obtained by the method of least squares and used for estimating the value of the dependent variable y for a known value of the independent variable x .

Estimation of Regression Parameters

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be n ordered pains of corresponding values of the two correlated variables x and y . Let the line to be estimated on the basis of these given n pairs of values of x and y by applying the method of least squares is of the form :



Let the estimated values of y corresponding to the values $x_1, x_2, ..., x_n$ of X be $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ respectively. Again, the actual values of y corresponding to the values $x_1, x_2, ..., x_n$ are $y_1, y_2, ..., y_n$ respectively. Corresponding to $x = x_1$, the value of y as estimated by the line (1) is $\hat{y}_1 = a + bx_1$.

The estimated value of y corresponding to $x = x_2$ is $\hat{y}_2 = a + bx_2$. Similarly, the estimated value of y corresponding to $x = x_n$ is $\hat{y}_n = a + bx_n$. If we denote the errors of estimates by $e_1, e_2, ..., e_n$ respectively then we have,

$$e_1 = y_1 - \hat{y}_1 = y_1 - (a + bx_1) = y_1 - a - bx_1$$

$$e_2 = y_2 - \hat{y}_2 = y_2 - (a + bx_2) = y_2 - a - bx_2$$

Similarly, $e_n = y_n - \hat{y}_n = y_n - (a + bx_n) = y_n - a - bx_n$

For a point above the line, the error of estimate will positive and for a point below the line, the error of estimate will be negative. Now the values of the two parameters a and b of the line (1) are to be determined in such a way that the sum of the squares of the errors is the least.

i.e. $\sum_{i-1}^{n} e^2{}_i = \sum_{i-1}^{n} (y_i - a - bx_i)^2$ is the least.

or $\Sigma e^2 = \Sigma (y - a - bx)^2$ is the least.

By applying the principle of least squares of Differential Calculus we find that if we put the values of a and b obtained by solving the following two equations in $y = a + bx$ then Σe^2 will the least with respect to this line.

$$\Sigma y = na + b\Sigma x \qquad -(2)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad -(3)$$

Equations (2) and (3) are called normal equations. From the given values of x and y we get Σx, Σy, Σx^2 and Σxy. Putting these values in (2) and (3) and solving these two equations simerltaneously for a and b we get the values of a and b. Putting these values of a and b in (1) we get the regression equation of y on x.

Dividing both sides of (2) by n we get,

$\bar{y} = a + b\bar{x}$ – (4)

Now, subtracting (4) from (1) we get,

$y - \bar{y} = b(x - \bar{x})$ – (5)

Again, multiplying (2) by Σx and (3) by n we get

$(\Sigma x)(\Sigma y) = na(\Sigma x) + b(\Sigma x)^2$

and $n(\Sigma xy) = na(\Sigma x) + bn(\Sigma x^2)$

Subtracting, $(\Sigma x)(\Sigma y) - n(\Sigma xy) = b\left[(\Sigma x)^2 - n\Sigma x^2\right]$

or, $b = \dfrac{(\Sigma x)(\Sigma y) - n(\Sigma xy)}{(\Sigma x)^2 - n\Sigma x^2}$

or, $b = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$ –(6)

In equation (5), b is called the regression coefficient of y on x and it is denoted by b_{yx}. Thus, if we write b_{yx} for b in the equation (5).

We get, $y - \bar{y} = b_{yx}(x - \bar{x})$ –(7)

Where $b_{yx} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$ –(8)

Equation (7) is called the regression equation of y on x where b_{yx} is given by (8). b_{yx} can also be expressed as

$$b_{yx} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} = \frac{\Sigma d_x d_y - \dfrac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma d_x{}^2 - \dfrac{(\Sigma d_x)^2}{n}}$$ (Assumed mean formula)

where $d_x = x - A$, A = assumed mean x

$d_y = y - B$, B = assumed mean y

There may be cases when the variable x depends on y and we may take the regression line of x on y as

$$x = a' + b'y \quad -(9)$$

Unlike the minimization of the sum of the squares of the errors for obtaining the estimates of a' and b' we get the following normal equations in a' and b' :

$$\Sigma x = na' + b'\Sigma y$$

$$\Sigma xy = a'\Sigma y + b'\Sigma y^2$$

By solving the above normal equations simultaniously for a' and b' and putting these values in (9) we get the regression equation of x on y as below :

$$x - \bar{x} = b'(y - \bar{y})$$

or $x - \bar{x} = b_{xy}(y - \bar{y})$ (we write b_{xy} for b')

Where $b_{xy} = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$

$$= \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}} = \frac{\Sigma d_x \Sigma d_y - \dfrac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma d_y{}^2 - \dfrac{(\Sigma d_y)^2}{n}}$$ (Assumed mean formula)

Where $d_x = x - A$, A = assumed mean of x

$d_y = y - B$, B = assumed mean of y .

Note : (i) The two lines of regression must intersect at the point $(\bar{x}, \bar{y})$. That is why this point satifies both the regression lines.

(ii) The regression equation of x on y is used to obtain the average value of y corresponding to a given value of x. Similarly, the regression equation of y on x is used to obtain the average value of x corresponding to a given of y .

5.5 PROPERTIES OF REGRESSION COEFFICIENTS

Property 1 : The product of the two regression coefficients is equal to the square of the correlation coefficient.

or

The correlation coefficient is the geometric mean of the regression coefficients.

Proof : We know that,

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\frac{1}{n}\Sigma(x-\bar{x})^2}$$

$$= \frac{Cov(x,y)}{\sigma x^2} = \frac{Cov(x,y)}{\sigma x.\sigma y}.\frac{\sigma_y}{\sigma_x} = r = \frac{\sigma_y}{\sigma_x}$$

Similarly, we can have,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

Now, $b_{yx}.b_{xy} = r\frac{\phi_y}{\phi_x}.r\frac{\phi_x}{\phi_y}$

$$\Rightarrow b_{yx}.b_{xy} = r^2$$

This shows that the product of the two regression coefficients is equal to the square of the correlation coefficient.

Again, from $r^2 = b_{yx}.b_{xy}$ we have,

$$r = \pm\sqrt{b_{yx}.b_{xy}}$$

Which shows that correlation coefficient r is the geometric mean of the two regression coefficients.

Property 2 : Both the regression coefficients must have the same sign.

Proof : We know that,

$$b_{yx} = \frac{Cov(x,\ y)}{\sigma x^2}$$

and $$b_{xy} = \frac{Cov(x,\ y)}{\sigma y^2}$$

In these two expressions, we must get $\sigma x^2 > 0$ and $\sigma y^2 > 0$ as squared quantities can't be negative. Moreover, both expressions have the same numerator $Cov(x,y)$.

Therefore, the sign of the two regression coefficients depends on the same term $Cov(x,y)$. Thus, if $Cov(x,y)$ is positive, both the regression

coefficients would be positive. Otherwise if $Cov(x, y)$ is negative, both the regression coefficients would be negative.

Property 3 : The two regression coefficients are independent of the change of origin but are dependent on the change of scale.

Proof : Let u and υ be the new variables obtained by changing the origin and scale of the original variable x and y as follows :

$$u = \frac{x-a}{h},\ \upsilon = \frac{y-b}{k} \qquad -(1)$$

Where, a, b, h and k (>0) are constants. Since correlation coefficient is independent of the change of origin and scale, we have,

$$r_{xy} = r_{uv} \qquad -(2)$$

As per transformation we have,

$$\sigma_x = h\sigma_u \text{ and } \sigma_y = h\sigma_\upsilon$$

Now, $$b_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x} = r_{u\upsilon}\frac{k\sigma_\upsilon}{h\sigma_u} = \frac{k}{h}\left(r_{u\upsilon}\frac{k_\upsilon}{h_u}\right) \qquad -(3)$$

$$b_{xy} = r_{xy}\frac{\sigma_x}{\sigma_y} = r_{u\upsilon}\frac{k\sigma_u}{h\sigma_\upsilon} = \frac{h}{k}\left(r_{u\upsilon}\frac{k_u}{h_\upsilon}\right) \qquad -(4)$$

From (3) and (4) it is clear that the two regression coefficients are independent of the change of origin but are dependent on the change of scale.



CHECK YOUR PROGRESS

Q.1: Fill in the blanks :

(i) The two regressions coefficients are of _____ sign.

(ii) Regression analysis is used to study_____ between the variables.

(iii) If a regression coefficient is negative then the correlation between the variables would also be _____.

(iv) If one of regression coefficient is greater than unity, the other one must be _____ than unity.

5.6 CORRELATION VS. REGRESSION

- With the help of correlation one measures the covariation between two variables. In correlation neither variable may be termed as dependent or independent variable. Since correlation does not establish a relationship between the two variables as such one can't estimate the value of one variable corresponding to a given value of the other variable.

  Regression establishes a functional relationship between two variables and hence one can estimate the value of one variable corresponding to a given value of the other variable.

- With the help of correlation one can't study which variable is the cause and which variable is the effect.

  Regression establishes the cause and effect relationship between two variables. The variable that indicates cause is the independent variable and the variable that indicates effect is the dependent variable.

- Correlatuon coefficient between two variables x and y is always symmetric, i.e. $r_{xy} = r_{yx}$. But regression coefficient is not symmetric in general, i.e. $b_{xy} \neq b_{yx}$ in general.

- Correlation coefficient is a relative measure of the linear relationship between the two variables x and y and hence it is a pure number, i.e. a unitless number whatever may be the unit of measurement of x and y. This pure number always lies between –1 and 1. On the other hand, regression coefficient is an absolute measure of change in one variable corresponding to one unit change in the other variable. Moreover, if one of the regression coefficient is greater than unity, the other regression coefficient must be less than unity.

- While correlation coefficient is independent of change of origin and scale, regression coefficients are independent of change of origin only but not of scale.

Examples

1. The following results were worked out from scores in Statisties and Mathematics in a certain examination :

	Scores in Statistics (X)	Scores in Mathematics (Y)
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Karl Pearson's correlation coefficient between X and Y is 0.42.

Find both the regression lines. Use these lines to estimate the value of Y when X = 50 and the value of X when Y = 30.

Solution :

We are given,

$\overline{X} = 39.5,\ \overline{Y} = 47.5$

$\sigma_x = 10.8,\ \sigma_y = 17.8$

and r_{xy} = 0.42

Now, $b_{YX} = r_{xy}\frac{\sigma_y}{\sigma_x} = 0.42\frac{17.8}{10.8}$ =0.69

$$b_{XY} = r_{xy}\frac{\sigma_y}{\sigma_x} = 0.42\frac{10.8}{17.8} = 0.25$$

$\therefore$ Regression line of Y on X is

$$Y - \overline{Y} = b_{YX}\left(X - \overline{X}\right)$$

$$\Rightarrow Y - 47.5 = 0.69(X - 39.5)$$

$$\Rightarrow Y = 0.69X - 0.69 \times 39.5 + 47.5$$

$$\Rightarrow Y = 0.69X + 20.24 \quad -(1)$$

$\therefore$ Regression line of X on Y is

$$X - \overline{X} = b_{XY}\left(Y - \overline{Y}\right)$$

$$\Rightarrow X - 39.5 = 0.25(Y - 47.5)$$

$$\Rightarrow X = 0.25Y - 0.25 \times 47.5 + 39.5$$

$$\Rightarrow X = 0.25Y + 27.62 \quad -(2)$$

Again, the predicted value of Y when X = 50 is obtained from the regression line of Y on X and is given by,

Y = 0.69 x 50 + 20.24 = 54.74

and the predicted value of X when Y = 30 is obtained from the regression line of X on Y and is given by,

X = 0.25 x 30 + 27.62 = 35.12

Example 2 : Given the bivariate data

x :	2	4	5	6	8	11
y :	18	12	10	8	7	5

(i) Fit the regression line of Y on X and predict when X = 10.

(ii) Fit the regression line of X on Y and predict when Y = 8.5.

(iii) Calculate Karl Pearson's coefficient of correlation.

Solution : Let us prepare the following table :

x	$x-\bar{x}$ $= x-6$	$(x-\bar{x})^2$	y	$y-\bar{y}$ $= y-10$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
2	–4	16	18	8	64	–32
4	–2	4	12	2	4	–4
5	–1	1	10	0	0	0
6	0	0	8	–2	4	-
8	2	4	7	–3	9	–6
11	5	25	5	–5	25	–25
$\Sigma X = 36$		$\Sigma(X-\bar{X})^2 = 50$	$\Sigma Y = 60$		$\Sigma(Y-\bar{Y})^2 = 106$	$\Sigma(X-\bar{X})(Y-\bar{Y}) = -67$

Now, $\bar{X} = \frac{\Sigma X}{n} = \frac{36}{6} = 6$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{60}{6} = 10$$

$$\therefore\ b_{YX} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$$

$$= \frac{-67}{50} = -1.34$$

$$b_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(Y-\bar{Y})^2}$$

$$= \frac{-67}{106} = -0.53$$

(i) The regression equation of Y on X is,

$$Y-\bar{Y} = b_{YX}(X-\bar{X})$$

$$\Rightarrow Y-10 = -1.34(X-6)$$

$$\Rightarrow Y = 1.34X + 1.34 \times 6 + 10$$

$$\Rightarrow Y = -1.34X + 18.04$$

When X = 10, the predicted value of y would be

y = –1.34 x 10 + 18.04

= 4.64

(ii) The regression equation of x on y is

$$X - \bar{X} = b_{XY}\left(Y - \bar{Y}\right)$$

$$\Rightarrow X - 6 = -0.63(Y - 10)$$

$$\Rightarrow X = -0.63Y + 0.63 \times 10 + 6$$

$$\Rightarrow X = -0.63Y + 12.3$$

When, y = 8.5 the predicted value of x would be,

x = –0.63 x 8.5 + 12.3

= 6.945

(iii) We know that,

$$r_{xy} = \pm\sqrt{b_{yx} . b_{xy}}$$

$$= -\sqrt{(-1.34) \times (-0.63)}$$

$$= -\sqrt{0.8442}$$

= –0.92 (approx)

Example 3 : Given below are the data for heights of some fathers and sons. Find out the two lines of regression. Calculate the expected height of the son when the height of the father is 67.5 inches.

Heights of father : (in inches)	65	66	67	67	68	69	71	73
Heights of son : (in inches)	67	68	64	58	72	70	69	70

Solution : Let, X denots the height of father and Y denotes the height of son.

Let us prepare the following table :

x	d_x=x-A =x-68	y	d_y=y-B =y-67	d^2x	d^2y	d_xd_y
65	–3	67	0	9	0	0
66	–2	68	1	4	1	–2
67	–1	64	–3	1	9	3
67	–1	58	–9	1	81	9
68	0	72	5	0	25	0
69	1	70	3	1	9	3
71	3	69	2	9	4	6
73	5	70	3	25	9	15
$\Sigma x = 546$	$\Sigma d_x = 2$	$\Sigma y = 538$	$\Sigma d_y = 2$	$\Sigma d^2{}_y = 50$	$\Sigma d^2{}_y = 138$	$\Sigma d_y d_y = 34$

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{546}{8} = 68.25$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{538}{8} = 67.25$$

Now, $$b_{yx} = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma d_x{}^2 - \frac{(\Sigma d_x)^2}{n}}$$

$$= \frac{34 - \frac{2 \times 2}{8}}{50 - \frac{2^2}{8}}$$

$$= \frac{34 - 0.5}{50 - 0.5} = \frac{33.5}{49.5} = 0.68$$

$$b_{xy} = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{n}}{\Sigma d_y{}^2 - \frac{(\Sigma d_y)^2}{n}}$$

$$= \frac{33.5}{138 - \frac{2^2}{8}} = \frac{33.5}{137.5} = 0.24$$

$\therefore$ The regression line of y on x is,

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\Rightarrow y - 67.25 = -0.68(x - 68.25)$$

$$\Rightarrow y = -0.68x - 0.68 \times 68.25 + 67.25$$

$$\Rightarrow y = -0.68x + 20.84$$

Again, the regression equation of x on y is,

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$\Rightarrow x - 68.25 = -0.24(y - 67.25$$

$$\Rightarrow x = 0.24y - 0.24 \times 67.25 + 68.25$$

$$\Rightarrow x = 0.24y + 52.11$$

The expected height of son when the height of father is 67.5 inches is to be obtained from the regression equation of y on x. Thus when x = 67.5 inches, the expected height of son, i.e. y = (0.68 x 67.5 + 20.84) inches = 66.74 inches.

Example 4 : For the variable x and y, the regression equations are given as 7x – 3y – 18 = 0 and 4x – y – 11 = 0

(i) Find the arithmetic means of x and y.

(ii) Identify the regression equation of y on x.

(iii) Compute the correlation coefficient between x and y.

(iv) Given the variances of x is 9, find the S.D. of y.

Solution :

(i) Since the two lines of regression intusect at the point $(\bar{x}, \bar{y})$, replacing x and y by $\bar{y}$ respectivelt in the given regression equations, we get,

$$7\bar{x} - 3\bar{y} - 18 = 0$$

and $$4\bar{x} - \bar{y} - 11 = 0$$

Solving these two equation, we get $\bar{x} = 3$ and $\bar{y} = 1$. Thus the arithmetic means of x and y is given by 3 and 1 respectively.

(ii) Let us assume that 7x – 3y – 18 = 0 represents the regression line of y on x and 4x – y – 11 = 0 represents the regression line of x on y.

Now, 7x – 3y – 8 = 0

$$\Rightarrow y = \frac{7}{3}x - 6$$

$$\therefore b_{yx} = \frac{7}{3}$$

Again 4x – y – 11 = 0

$$\Rightarrow x = \frac{1}{4}y + \frac{11}{4}$$

$$\therefore\ b_{xy} = \frac{1}{4}$$

Thus $r^2 = b_{yx} \times b_{xy} = \frac{7}{3} \times \frac{1}{4}$

$$= \frac{7}{12} < 1$$

Since $|r| \leq 1 \Rightarrow r^2 \leq 1$, our assumptions are correct. Thus, 7x – 3y – 18 = 0 truly represents the regression line of y on x.

(iii) Since $r^2 = \frac{7}{12}$

$$\therefore\ r = \sqrt{\frac{7}{12}} = 0.7638$$

(iv) We know that,

$$b_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x}$$

$$\Rightarrow \frac{7}{3} = 0.7638 \times \frac{\sigma_y}{3} \qquad [\because\ Var(x) = \sigma_x{}^2 = 9]$$

$$\Rightarrow \sigma_y = \frac{7}{0.7638} = 9.1647$$

Example 5 : For a bivariate data, you are given the following information :

$\Sigma(x-58) = 46$ $\Sigma(x-58)^2 = 3086$

$\Sigma(y-68) = 9$ $\Sigma(y-68)^2 = 483$

$\Sigma(x-58)(y-68) = 1095,\ \bar{x} = 58,\ \bar{y} = 68$

Number of pairs of observations = 7. You are required to determains.

(i) the two regression equations and

(ii) the coefficient of correlation between x and y.

Solution : We are given, $\bar{x} = 58,\ \bar{y} = 68$

$$\Sigma(x - \bar{x}) = \Sigma(x - 58) = 46$$

$$\Sigma(x - \bar{x})^2 = \Sigma(x - 58)^2 = 3086$$

$$\Sigma(y - \bar{y}) = \Sigma(y - 68) = 9$$

$$\Sigma(y - \bar{y})^2 = \Sigma(y - 68)^2 = 483$$

$$\Sigma(x-\bar{x})(y-\bar{y}) = \Sigma(x-58)(y-68) = 1095$$

Now, $b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{1095}{3086} = 0.35$

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{1095}{483} = 2.27$$

(ii) $\therefore$ The regression line of y on x is

$$y-\bar{y} = b_{yx}(x-\bar{x})$$

$$\Rightarrow y-68 = 0.35(x-58)$$

$$\Rightarrow y = 0.35x - 0.35\times 58 + 68$$

$$\Rightarrow y = 0.35x + 47.7$$

The regression line of x on y is

$$x-\bar{x} = b_{xy}(y-\bar{y})$$

$$\Rightarrow x-58 = 2.27(y-68)$$

$$\Rightarrow x = 2.27y - 2.27\times 68 + 58$$

$$\Rightarrow x = 2.27y - 96.36$$

(ii) The coefficient of correlation

$$r_{xy} = \pm\sqrt{b_{yx}\,.\,b_{xy}}$$

$$= +\sqrt{0.35\times 2.27} = 0.89$$

Example : Find out the regression coefficients Y on X, X on Y and correlation coefficient between X and Y on the basis of the following data :

$\sum xy = 350,\ \sum x = 50,\ \sum y = 60,\ n = 10,$

Variance of x=4 and Variance of Y=9

Solution : Regression coefficient of Y on X is given by

$$b_{yx} = \frac{Cov(x,y)}{\sigma_x^{\,2}}$$

$$= \frac{\frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{Variance\ of\ X}$$

$$= \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{Variance\ of\ X}$$

$$= \frac{\frac{1}{n}\sum xy - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{Variance\ of\ X}$$

$$= \frac{\frac{350}{10} - \frac{50}{10} \times \frac{60}{10}}{4} = 1.25$$

Regression Coefficient of X on Y is given by

$$b_{xy} = \frac{Cov(x, y)}{\sigma_y^{\,2}}$$

$$= \frac{\frac{1}{n}\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{Variance\ of\ Y}$$

$$= \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{Variance\ of\ Y}$$

$$= \frac{\frac{1}{n}\sum xy - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)}{Variance\ of\ Y}$$

$$= \frac{\frac{350}{10} - \frac{50}{10} \times \frac{60}{10}}{9}$$

$$= \frac{\frac{350}{10} - \frac{50}{10} \times \frac{60}{10}}{9}$$

Hence, Coefficient of correlation between X and Y is given by

$$r_{xy} = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

$$= \pm\sqrt{1.25 \times 0.55}$$

$$= 0.83$$

5.7 MULTIPLE CHOICE QUESTIONS

Q 1: The regression line of y on x is derived by

(a) the minimisation of vertical distances in the scatter diagram.

(b) the minimisation of horizontal distances in the scatter diagram.

(c) both (a) and (b)

(d) (a) or (b)

Ans : (a)

Q 2: The two lines of regression become identical when

(a) r = 1 (b) r = –1

(c) r = 0 (d) (a) or (b)

Ans : (d)

Q 3: What are the limits of the two regression coefficients ?

(a) No limit (b) Must be positive

(c) One positive and the other negative

(d) Product of the regression coefficients must be numerically less than unity.

Ans : (d)

Q 4: The regression coefficients remain unchanged due to

(a) shift of origin (b) shift of scale

(c) both (a) and (b) (d) (a) or (b)

Ans : (a)

Q 5: The two line of regression intersect at the point

(a) (0, 0) (b) (1, 1)

(c) (–1, –1) (d) $(\bar{x}, \bar{y})$

Ans : (d)

Q 6: If x and y are two variables, there can be at most

(a) one regression line

(b) two regression line

(c) three regression line

(d) an infinite number of regression lines

Ans : (b)

Q 7: The errors in case of regression equations are

(a) positive (b) negative

(c) zero (d) All these

Ans : (d)

Q 8: The regression lines are perpendicular to each other if r is equal to

(a) 0 (b) +1

(c) –1 (d) ± 1

Ans : (a)

Q 9: The regression coefficients are zero if r is equal to

(a) 2 (b) –1

(c) 1 (d) 0

Ans : (d)

Q 10: In the regression line Y = a + bx, b is called the

(a) slope of the line

(b) intercept of the line

(c) neither (a) nor (b)

(d) both (a) and (b)

Ans : (a)

5.8 EXERCISES

Q 1: What do you mean by linear regression? Why there are two regression lines ? When do these become identical ?

Q 2: Obtain the equations of the two lines of regression for a bivariate distribution.

Q 3: If the regression coefficient of y on x, the coefficient of correlation between x and y and variance of y are $-3/4$, $-\sqrt{3/4}$ and 4 respectively, what is the variance of x ?

Ans : $16/3$

Q 4: The following data are about sale and advertisement expenditures of a firm :

	Sale	Advertisement expenditure (in crores of Rs.)
Mean	40	6
S. D.	10	1.5

Coefficient of correlation r = 0.9

(a) Estimate the likely sale for a proposed advertisement expenditure of Rs. 10 crores.

(b) What would be the advertisement expenditure if the firm fixes a sale target of 60 crores of rupees ?

Q 5: The following results relate to bivariate data on (x, y) :

$\Sigma x = 142,$ $\Sigma y = 166$

$\Sigma xy = 2434.$ $\Sigma x^2 = 2085$

$\Sigma y^2 = 2897$

Determine the correlation coefficient between x and y. Estimate y at x = 20 using linear regression.

(Ans : r = 0.78, y = 23.56)

Q 6: Given the variance of x = 9. The regression equations are

8x – 10y + 66 = 0 and 40x – 18y – 214 = 0

What were

(a) The means of x and y

(b) The coefficient of correlation between x and y

(c) The standard deviation of y

(Ans : (a) $\bar{x} = 13, \bar{y} = 17$

(b) r = 0.6

(c) σ_y =4)

Q 7: Find the two regression lines from the following data :

x :	1	2	3	4	5
y :	2	3	5	4	6

Estimate y if x = 2.5

(Ans : y = 0.9x + 1.3, x = 0.9y – 0.6, y = 3.55)

Q 8: The following data were obtained from two variable x and y :

$\bar{x} = 20,\ \bar{y} = 15,\ \sigma_x = 4,\ \sigma_y = 3,\ r = 0.7$

Obtain the two regression equations and find the most likely value of y when x = 24.

(Ans : y = 0.53x + 4.5, x = 0.93y + 6, y = 17.22)

Q 9: Assume that the two statistical variables x and y follow the regression equations

y = 2x + 2

and x = ay + b

If the correlation coefficient between x and y is 0.93, find the value of a.

[Ans : a = 0.43 (approx)]



5.9 LET US SUM UP

- Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.
- The two equations which are estimated on the basis of the given values of the two variables to estimate the most likely values of one variable for

specified values of the other are called regression equations.

- If the points of the scatter diagram of two correlated variables clusters around a straight line then we say that there is linear regreassion between the two variables.
- When only two variables are associated with regression analysis then we call it single regression.
- The product of the two regression coefficients is equal to the square of the correlation coefficient or The correlation coefficient is the geometric mean of the regression coefficients.
- Both the regression coefficients must have the same sign.
- The two regression coefficients are independent of the change of origin but are dependent on the change of scale.
- With the help of correlation one measures the covariation between two variables.
- In correlation neither variable may be termed as dependent or independent variable.
- Since correlation does not establish a reltionship between the two variables as such one can't estimate the value of one variable corresponding to a given value of the other variable.
- Regression establishes a functional relationship between two variables and hence one can estimate the value of one variable corresponding to a given value of the other variable.
- With the help of correlation one can't studt which variable is the cause and which variable is the effect.
- Regression establishes the cause and effect relationship between two variables.
- The variable that indecates cause is the independent variable and the variable that indicates effect is the dependent variable.
- Correlatuon coefficient between two variables x and y is always symmetric, i.e. $r_{xy} = r_{yx}$. But regression coefficient is not symmetric in general, i.e. $b_{xy} \neq b_{yx}$ in general.
- Correlation coefficient is a relative measure of the linear relationship between the two variables x and y and hence it is a pure number, i.e. a

unitless number whatever may be the unit of measurement of x and y. This pure number always lies between –1 and 1.

- Regression coefficient is an absolute measure of change in one variable corresponding to one unit change in the other variable. Moreover, if one of the regression coefficient is greater than unity, the other regression coefficient must be less than unity.
- While correlation coefficient is independent of change of origin and scale, regression coefficients are independent of change of origin only but not of scale.



5.10 FURTHER READING

1) Gujrati, D (1995); Basic Econometrics (third edition), McGraw-Hill, Inc; Singapore
2) Gun, A M; Gupta, M K & Dasgupta, B (2002); Fundamentals of Statistics, Volume One; World Press Private Ltd. Kolkata
3) Gupta, S C (1994); Fundamentals of Statistics; Himalayan Publishing House, New Delhi.



5.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q. No. 1: (i) same (ii) dependence
(iii) negative (iv) less



5.12 MODEL QUESTIONS

A. SHORT QUESTIONS

Q 1: What is meant by regression? What are its types?

Q 2: Distinguish between linear and non-linear regression.

Q 3: State the limits of two regression coefficient.

Q 4: What is a two variable regression model?

B. LONG QUESTIONS

Q 1: State the properties of regression coefficients and prove them.

Q 2: Discuss the relationship between correlation and regression.

Q 3: Given below the information about the capital employed and profit earned by a company over the last twenty five years :

	Mean	S.D.
Capital employed ('0000 Rs.)	62	5
Profit earned ('000 Rs.)	25	6

Correlation coefficient between capital and profit = 0.92. Find the sum of the regression coefficients for the above data.

(Ans : 1.871)

*** ***** ***

UNIT 6: PROBABILITY : BASIC IDEAS

UNIT STRUCTURE

6.1 Learning Objectives

6.2 Introduction

6.3 Definition of the Basic Terms

6.4 Classical or Mathematical or Apriori Definition of Probability

6.5 Addition Theorems or Theorems on Total Probability

6.6 The Theorem of Compound Probability or the Multiplicative Law of Probability

6.7 Baye's Theorem (Concept only)

6.8 Exercises

6.9 Let Us Sum Up

6.10 Further Reading

6.11 Answers to Check Your Progress

6.12 Model Questions

6.1 LEARNING OBJECTIVES

After studying this chapter, you will be able to :

- discuss the basic concept of probability
- define the meanings of the terms related to probability
- put foward the different rules and their applications
- explain the different computational techniques of solving practical problems.

6.2 INTRODUCTION

The concept of probability originated from the analysis of the games of chance in the 17th century. In recent time, probability has developed itself into a full fledged subject and become an integral part of Statistics. The theopry of probability is a study of statistical or Random experiments. It is the back bone of Statistical inference and Decision Theory that are essential tools of the analysis of most of the modern business and economic problems.

The broad divisions of probability are Subjective Probability and Objective Probability. Subjective Probability is basically dependent on personal judgement and experience and, as such, it may be influenced by the personal belief, attitude and bias of the person applying it. However, in the field of uncertainity, this would be quite helpful and it is being applied in the area of decision making management. This Subjective Probability is beyond the scope of our present discussion. We ar going to discuss Objective Probability in the remaining sections.

6.3 DEFINITION OF THE BASIC TERMS

In this section, we shall discuss the basic definitions of a few terms relationg to the concept of probalility.

Random Experiment: If the result or outcome of an experiment conducted again and again under the same conditions may be any one of the possible outcomes then the experiment is called a random experiment. For example, if a coin is tossed again and again under the same conditions then we get two outcomes - Head (H) and Tail (T). It is impossible to say in advance whether a Head or a Tail would turn up when we toss the coin once. Thus, tossing a coin is an example of a random experiment.

Trial: The task of performing a random experiment is called a trial. For example, if a coin is tossed then it would be a trial.

Event: The outcome or result of a trial is called an event. For example, if a coin is tossed and we get a head it would be called an event.

Equally likely events: Two or more events are said to be equally likely if neither of them occurs in preference to the other. For example, if an unbiased coin is tossed, the two possible outcomes or events, a head or a tail are equally likely.

Mutually Exclusive events: Two or more events are said to be mutually exclusive if they are such that when one occurs, other (others) cannot occur. For example, the two possible outcomes or events of toss of a coin are mutually exclusive.

Exhaustive events: The events associated with a random experiment are said to be exhaustive if they include all possible outcomes of the experiment. Thus, if a set of events are exhaustive then any one of

them is sure to occur. For example, the two events Head and Tail, when a coin is tossed once, are exhaustive as no other event except these two can occur.

Favourable cases: The outcomes or events of a random experiment which imply the occurrence of an event are called the favourable cases to the event. For example, the number of favourable cases to the event of getting an even number in throwing a dice is 3, since there are 3 even numbers 2, 4, and 6 in between 1 and 6.

Sample Space: The set of all possible outcomes of a random experiment is called a Sample Space and generally it is denoted by the symbol S or Ω. For example, if a dice is thrown then the Sample Space would be: S = {1,2,3,4,5,6}

6.4 CLASSICAL OR MATHEMATICAL OR APRIORI DEFINITION OF PROBABILITY

If n is the number of equally likely, mutually exclusive and exhaustive outcomes of a random experiment out of which m outcomes are favourable to the occurrence of an event A, then the probability that A occurs, denoted by P(A), is given by:

$$P(A) = \frac{Number\ of\ outcomes\ favourable\ to\ A}{Number\ of\ exhaustive\ outcomes\ or\ total\ number\ of\ probable\ outcomes}$$

$$= \frac{m}{n}$$

The probability of non-occurrence of the event A is denoted by $P(\overline{A})$ or $P(A^c)$ and is given by

$$P(\overline{A}) = \frac{Number\ of\ outcomes\ unfavourable\ to\ A}{Total\ number\ of\ probable\ outcomes}$$

$$= \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n}$$

$$= 1 - \frac{m}{n}$$

Now, $P(\overline{A}) = 1 - \frac{m}{n}$

$\Rightarrow P(\overline{A}) = 1 - P(A)$

$\Rightarrow P(A) + P(\overline{A}) = 1$

Note 1: If the number of favourable cases, m = n i.e. all the probable cases are favourable, then probability of the event A becomes 1, i.e. P(A) = 1. In this case A is called a certain or sure event.

If the number of favourable cases, m = 0, i.e. all the probable cases are unfavourable, then probability of the event A becomes 0, i.e. P(A) = 0. In this case A is called an impossible event.

Since the value of m lies between 0 and n, the probability of event A lies between 0 and 1, i.e. $0 \leq P(A) \leq 1$.

Note 2: The classical definition of probability has the following limitations:

(a) It is applicable only when the total number of events is finite.

(b) It can be used only when the events are equally likely or equi - probable. This assumption is made well before the experiment is performed.

(c) This definition has only a limited field of application like coin tossing, dice throwing, drawing cards etc. where the possible events are known well in advance. In the field of uncertainity or where no prior knowledge is provided, this definition is inapplicable.

Note 3: The ratio of number of favourable events is known as odds in favour of the event A and its inverse ratio is known as odds against the event A. i.e.

odds in favour of A = $\frac{m}{n-m}$

and odds against A = $\frac{n-m}{m}$



CHECK YOUR PROGRESS

Q 1: State True or False

(i) The probability of an impossible event is always zero.

(ii) If two events are independent, then they will also be mutually exclusive.

(iii) The probability of occurence of an event cannot be a negative number.

Example 1: A coin is tossed 3 times. What is the probability of getting : i) 2 heads ii) atleast 2 heads

Solution: When a coin is tossed 3 times, we shall get the following probable outcomes:

HHH, HHT, HTH, HTT, THH, THT, TTH, TTT i.e. 8 outcomes

(i) Out of these 8 outcomes, 2 heads occur in three cases namely HHT, HTH and THH. If A denotes the event of getting 2 heads, then above three cases are favourable for the event A.

$$\therefore P(A) = \frac{3}{8} = 0.375$$

(ii) Let B denotes the event of getting atleast 2 heads. The favourable cases for the event B are,

HHT, HTH, THH, HHH i.e. 4 cases

$$\therefore P(B) = \frac{4}{8} = 0.50$$

Example 2: Two dice are thrown simultaneously. Find the probability that the sum of points on the two dice would be 7 or more.

Solution: If two dice are thrown, then the probable outcomes would be

(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6)
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)
i.e. 36 cases

Now, the cases where the sum of points is 7 or more are -

(1,6), (2,5), (2,6), (3,4), (3,5)
(3,6), (4,3), (4,4), (4,5), (4,6)
(5,2), (5,3), (5,4), (5,5), (5,6)
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)
i.e. 21 cases

$\therefore$ Probability that the sum of points on the two dice is 7 or more

$$= \frac{21}{36} = \frac{7}{12}$$

6.5 ADDITION THEOREMS OR THEOREMS OF TOTAL PROBABILITY

Theorem 1:

Statement: If $A_1, A_2, \ldots\ldots\ldots\ldots, A_k$ are k mutually exclusive and exhaustive events then the probability that any one of them occurs is equal to the sum of the probabilities of occurrences of these events separately. In symbol,

$P(A_1+A_2+\ldots\ldots\ldots+A_k)$ or $P(A_1 \cup A_2 \cup \ldots\ldots\ldots\ldots\ldots \cup A_k) = P(A_1)+P(A_2)+\ldots\ldots\ldots\ldots+P(A_k)$

Theorem 2:

Statement: If A and B are two events which are not mutually exclusive then the probability that atleast one event will occur i.e. P(A+B) or $P(A \cup B)$ is given by

$P(A+B)$ or $P(A \cup B) = P(A)+P(B) - P(AB)$, where $P(AB) = P(A \cap B)$

Example 3: If two dice are thrown what is the probability that a sum of nine points or eleven points will be obtained?

Solution: Let A: the event that the sum of points is nine

B: the event that the sum of points is eleven

Now, when two dices are thrown, we may get 36 different outcomes. Hence the total number of all possible cases = 36.

Again, the cases favourable for the event A are: (6,3), (5,4), (4,5), (3,6), i.e. 4 cases.

The cases favourable for the event B are: (6,5), (5,6) i.e. 2 cases.

$$\therefore P(A) = \frac{4}{36}$$

and, $$P(B) = \frac{2}{36}$$

Here, the events A and B are mutually exclusive.

$\therefore$ P (that a sum of nine points or eleven points)

$= P(A \cup B)$

$= P(A) + P(B)$

$= \frac{4}{36} + \frac{2}{36}$

$= \frac{6}{36} = \frac{1}{6}$

Example 4: There are 20 marbles in an urn which are numbered from 1 to 20. A marble is drawn at random. Find the probability that the number on the marble is a multiple of either 2 or 5.

Solution:

Let, A: the event that the number on the marble is a multiple of 2

B: the event that the number on the marble is a multiple of 5.

The cases favourable for the event A are: 2,4,6,8,10,12,14,16,18,20; i.e. 10 cases

The cases favourable for the event B are: 5, 10, 15, 20; i.e. 4 cases. Again, the cases favourable for the vent both A and B, i.e. favourable for $A \cap B$ are: 10,20; i.e. 2 cases

Here, the events A and B are not mutually exclusive. Therefore, we have

P (that the number is a multiple of either 2 or 5)

$= P(A \cup B) = P(A) + P(B) - P(AB)$

$= \frac{10}{20} + \frac{4}{20} - \frac{2}{20} = \frac{12}{20} = \frac{3}{5}$

6.6 THE THEOREM OF COMPOUND PROBABILITY OR THE MULTIPLICATIVE LAW OF PROBABILITY

Statement: For any two events A and B, the probability that A and B occur simultaneously is given by the product of the unconditional probability of A and the conditional probability of B given that A has already occured, i.e.

$P(A \cap B)$ or $P(AB) = P(A) \cdot P(B/A)$, provided $P(A) > 0$

Alternatively, we may write

$P(A \cap B)$ or $P(AB) = P(B) \cdot P(A/B)$, provided $P(B) > 0$

Note:

1. If the events A and B independent, then P(B/A) = P(B) and consequently for two independent events A and B the multiplicative law becomes P(AB) = P(A) . P(B). Thus we may also define independent event as

Two events A and B are independent if

$$P(AB) = P(A).P(B)$$

2. From the theorem of conditional probability, we have,

$$P(AB) = P(A).P(B/A)$$

Example 5: An urn contains 6 white and 8 black balls. Two successive draws of one ball are made. Find the probability that the first drawn ball is white and the second drawn ball is black if

(i) the first drawn ball is replaced before the second draw.

(ii) the first drawn ball is not replaced.

Solution:

Let A: the event of drawing a white ball in the first draw

B: the event of drawing a black ball in the second draw.

Case (i): The first drawn ball is replaced

If the first drawn ball is replaced before the second draw, the event A will have no effect on the occurrence of the event B as the total number of balls will remain the same just before the second draw. Hence, the events A and B are independent of each other.

$$\therefore P(AB) = P(A).P(B)$$

$$= \frac{6}{14} \times \frac{8}{14} = \frac{48}{196}$$

$$= \frac{6}{14} \times \frac{8}{14} = \frac{48}{196}$$

Case (ii): The first drawn ball is not replaced

If the first drawn ball is not replaced, the event A will affect the occurrence of the event B. Hence, we have,

$$P(AB) = P(A) \cdot P(B/A)$$

$$= \frac{6}{14} \times \frac{8}{13} = \frac{48}{182}$$

Certain Other Theorems Relating to Probability

Theorem 1: If A and B are independent events then A and $\overline{B}$ are also independent events.

Proof: The event A can be expressed as follows:

$$A = (A \cap \overline{B}) \cup (A \cap B)$$

Here, $A \cap \overline{B}$ and $A \cap B$ are disjoint or mutually exclusive events.

Hence, by Axiomatic definition of probability, we get,

$$P(A) = P(A \cap \overline{B}) + P(A \cap B)$$

$$\Rightarrow P(A \cap \overline{B}) = P(A) - P(A \cap B)$$

$$= P(A) - P(A).P(B) \quad [\because \text{A and B are independent}]$$

$$= P(A)[1 - P(B)]$$

$$= P(A)P(\overline{B})$$

$\therefore$ A and $\overline{B}$ are independent

Theorem 2: If A and B are independent events then $\overline{A}$ and $\overline{B}$ are also independent events.

Proof: We are given that

$$P(A \cap B) = P(A).P(B)$$

Now, $P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B})$

$$= 1 - P(A \cup B)$$

$$= 1 - [P(A) + P(B) - P(A \cap B)]$$

$$= 1 - [P(A) + P(B) - P(A).P(B)] \quad (\because \text{A and B are independent})$$

$$= 1 - P(A) - P(B) + P(A).P(B)$$

$$= [1 - P(B)] - P(A)[1 - P(B)]$$

$$= [1 - P(B)][1 - P(A)]$$

$$= P(\overline{A})P(\overline{B})$$

Hence $\overline{A}$ and $\overline{B}$ are independent.

Example 6: If A and B are two events such that P(A) = $\frac{2}{3}$, $P(\overline{A} \cap B) = \frac{1}{6}$ and $P(A \cap B) = \frac{1}{3}$, find $P(B), P(A \cup B), P(A/B), P(B/A), P(\overline{A} \cup B), P(\overline{A} \cap \overline{B})$ and $P(\overline{B})$. Also examine whether the events A and B are: (a) Equally likely, (b) Exhaustive, (c) Mutually exclusive and (d) Independent.

Solution:

We are given, P(A) = $\frac{2}{3}$, $P(\overline{A} \cap B) = \frac{1}{6}$, $P(A \cap B) = \frac{1}{3}$

$$\therefore P(B) = P(\overline{A} \cap B) + P(A \cap B)$$

$$= \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

Now, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{2}{3} + \frac{1}{2} - \frac{1}{3} = \frac{5}{6}$$

Again, $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3} \times \frac{2}{1} = \frac{2}{3}$

and, $P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1}{3} \times \frac{3}{2} = \frac{1}{2}$

Now, $P(\overline{A} \cup B) = P(\overline{A}) + P(B) - P(\overline{A} \cap B)$

$$= [1 - P(A)] + P(B) - P(\overline{A} \cap B)$$

$$= 1 - \frac{2}{3} + \frac{1}{2} - \frac{1}{6}$$

$$= \frac{2}{3}$$

$$P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B)$$

$$= 1 - \frac{5}{6} = \frac{1}{6}$$

$P(\overline{B})$ = 1 - P(B) = 1- $\frac{1}{2}$ = $\frac{1}{2}$

(a) Since $P(A) \neq P(B)$, A and B are not equally likely events.

(b) Since $P(A \cup B) \neq 1$, A and B are not exhaustive events.

(c) Since $P(A \cap B) \neq 0$, A and B are not mutually exclusive events.

(d) Since, $P(A).P(B) = P(A \cap B)$, A and B are independent events.

Example 7: A bag contains 5 red and 3 black balls and second bag contains 4 red and 5 black balls.

(a) If one ball is selected at random from each bag, what is the probability that both of them are of same colour?

(b) If a bag is selected at random and two balls are drawn from it, what is the probability that they are of (i) same colour, (ii) different colour?

Solution:

(a) Required probability = [Probability that ball from both bags are red] + [Probability that balls from both bags are black]

$$= \frac{5}{8}\times\frac{4}{9} + \frac{3}{8}\times\frac{5}{9}$$

$$= \frac{35}{72}$$

(b) Let A be the event that first bag is drawn, so that $\overline{A}$ denotes the event that second bag is drawn. Since the two events are equally likely, mutually exclusive and exhaustive, we have

$$P(A) = P(\overline{A}) = \frac{1}{2}$$

(i) Let R be the event that two drawn balls are red and B be the event that they are black.

$\therefore$ The required probability

$= [P(A \cap R) + P(A \cap B)] + [P(\overline{A} \cap R) + P(\overline{A} \cap B)]$

$= P(A)P(R/A) + P(A)P(B/A) + P(\overline{A})P(R/\overline{A}) + P(\overline{A})P(B/\overline{A})$

$= P(A)[P(R/A) + P(B/A)] + P(\overline{A})[P(R/\overline{A}) + P(B/\overline{A})]$

$$= \frac{1}{2}\left[\frac{{}^5c_2}{{}^8c_2} + \frac{{}^3c_2}{{}^8c_2}\right] + \frac{1}{2}\left[\frac{{}^4c_2}{{}^9c_2} + \frac{{}^5c_2}{{}^9c_2}\right]$$

$$= \frac{1}{2}\left[\frac{10+3}{28}\right] + \frac{1}{2}\left[\frac{6+10}{36}\right] = \frac{229}{504}$$

(ii) Let c denotes the event that the drawn balls are of different colours.

$\therefore$ The required probability

$$P(C) = P(A \cap C) + P(\overline{A} \cap C)$$

$$= P(A)P(C/A) + P(\overline{A})P(C/\overline{A})$$

$$= \frac{1}{2}\cdot\frac{5\times 3}{{}^8c_2} + \frac{1}{2}\cdot\frac{4\times 5}{{}^9c_2}$$

$$= \frac{1}{2}\left[\frac{15}{28} + \frac{20}{36}\right] = \frac{275}{504}$$

Example 8: The odds against A solving a certain problem are 4 to 3 (or 4:3) and odds in favour of B solving the same problem are 7 to 5 (or 7:5). What is the probability that the problem is solved if they both try independently?

Solution:

[Note: If a is the number of cases favourable to an event A and $\propto$ is the number of cases favourable to its compliment event $\overline{A}$, then odds in favour of A are $a:\propto$ and odds against A are $\propto: a$. Obviously $P(A)=\frac{a}{a+\propto}$ and $P(\overline{A})=\frac{\propto}{a+\propto}$]

Let, E_1 denotes the event that A solves the problem

E_2 denotes the event that B solves the problem

Clearly E_1 and E_2 ar independent event.

We are given that, $P(E_1) = \frac{3}{4+3} = \frac{3}{7}$

$$\Rightarrow P(\overline{E}_1)=1-\frac{3}{7}=\frac{4}{7}$$

and, $P(E_2)=\frac{7}{7+5}=\frac{7}{12}$

$$\Rightarrow P(\overline{E}_2)=1-\frac{7}{12}=\frac{5}{12}$$

Now, P(atleast one can solve the problem) = 1- P(none can solve the problem)

$$= 1-P\left(\overline{E}1\cap\overline{E}2\right)$$
$$= 1-P\left(\overline{E}_1\right).P\left(\overline{E}_2\right)$$
$$= 1-\frac{4}{7}\times\frac{5}{12}=\frac{16}{21}$$

Example 9: A sum is given to five students A, B, C, D and E. Their respective chances of solving it are $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{4}$ and $\frac{1}{5}$. What is the probability that atleast one of the students solves the sum?

Solution:

Let, E1: the event of solving the sum by A

E2: the event of solving the sum by B

E3: the event of solving the sum by C

E4: the event of solving the sum by D

E5: the event of solving the sum by E

We are given,

$$P(E_1)=\frac{1}{2},\ P(E_2)=\frac{1}{3},\ P(E_3)=\frac{1}{4},\ P(E_4)=\frac{1}{4},\ P(E_5)=\frac{1}{5}$$

Now, P(atleast one can solve the sum) = 1- P(None can solve the sum)

$$= 1-P\left(\overline{E}_1\cap\overline{E}_2\cap\overline{E}_3\cap\overline{E}_4\cap\overline{E}_5\right)$$

$$= 1-P\left(\overline{E}_1\right)P\left(\overline{E}_2\right)P\left(\overline{E}_3\right)P\left(\overline{E}_4\right)P\left(\overline{E}_5\right)$$

$$= 1-\left(1-\frac{1}{2}\right)\left(1-\frac{1}{3}\right)\left(1-\frac{1}{4}\right)\left(1-\frac{1}{5}\right)$$

$$= 1-\left(\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}\times\frac{3}{4}\times\frac{4}{5}\right)$$

$$= 1-\frac{3}{20}=\frac{17}{20}$$

Example 10: There are two urns containing 5 red and 6 white balls and 3 red and 7 white balls respectively. If two balls are drawn from the first urn without replacement and transferred to the second urn and then a draw of another two balls is made from it, what is the probability that both the balls drawn are red?

Solution:

Since two balls are transferred from the first urn containing 5 red and 6 white balls to the second urn containing 3 red and 7 white balls, we are to consider the following cases:

Case A: Both the balls transferred are red. In this case, the second urn contains 5 red and 7 white balls.

Case B: The two balls transferred are of different colours. Then the second urn contains 4 red and 8 white balls.

Case C: Both the balls transferred are white. Now the second urn contains 3 red and 7 white balls.

Let R denotes the event of drawing two red balls from the second urn.

$\therefore$ The required probability = $P\left(R\cap A\right)+\left(R\cap B\right)+P\left(R\cap C\right)$

$$= P(A)P(R/A)+P(B)P(R/B)+P(C)P(R/C)$$

$$= \frac{{}^5c_2}{{}^{11}c_2} \times \frac{{}^5c_2}{{}^{12}c_2} + \frac{{}^5c_1 \times {}^6c_1}{{}^{11}c_2} \times \frac{{}^4c_2}{{}^{12}c_2} + \frac{{}^6c_2}{{}^{11}c_2} \times \frac{{}^3c_2}{{}^{12}c_2}$$

$$= \frac{10}{55} \times \frac{10}{66} + \frac{30}{55} \times \frac{6}{66} + \frac{15}{55} \times \frac{3}{66}$$

$$= \frac{325}{66 \times 55} = \frac{65}{726}$$



CHECK YOUR PROGRESS

Q 2: Fill in the blanks

(i) Two events A and B are statistically independent when P(A)=_____.

(ii) The probability of getting and odd number in tossing a dice is_____.

(iii) If P(A/B)=P(A), then the events A and B are _____.

(iv) If $P(A \cap B) = 0.60$ and $P(A \cap B) = 0.70$ of two events A and B, then $P(A) + P(B) =$ _____.

6.7 BAYE'S THEOREM (CONCEPT ONLY)

This theorem is also known as theorem of posteriori probability. It was developed by Rev. Thomas Bayes.

The Baye's theorem is basically a technique of determining prior probabilities in the light of new information. The rule is stated below:

Let an event A can ocuur if any one of the set of exhaustive and mutually exclusive events B1, B2, ..., Bn occurs. The probabilities P(B1), P(B2), ..., P(Bn) and the conditional probabilities P(A/Bi), i = 1,2,....., n for A to ocuur are known. Then the conditional probability P(Bi/ A) when A has actually occured is given by:

$$P(B_i / A) = \frac{P(B_i).P(A / B_i)}{P(B_1).P(A / B_1) + P(B_2).P(A / B_2) + \ldots\ldots\ldots\ldots + P(B_n).P(A / B_n')}$$

$i = 1,2,\ldots\ldots,n$

6.8 EXERCISES

Multiple Choice Questions

Q 1: If $P(A \cap B) = 0$, then the two events A and B are

(a) Mutually exclusive (b) Exhaustive

(c) Equally likely (d) Independent

Ans: (a)

Q 2: If for two events A and B, $P(A \cup B) = 1$, then A and B are

(a) Mutually Exclusive events (b) Equally likely events

(c) Exhaustive events (d) Dependent events

Ans: (c)

Q 3: If P(A) = P(B), then the two events A and B are

(a) Independent (b) Dependent

(c) Equally Likely (d) Both (a) and (c)

Ans: (c)

Q 4: If for two events A and B, $P(A \cap B) \neq P(A) \times P(B)$, then the two events A and B are

(a) Independent (b) Dependent

(c) Not equally likely (d) Not exhaustive

Ans: (b)

Q 5: If $P(A/B) = P(A)$, then

(a) A is independent of B (b) B is independent of A

(c) B is dependent of A (d) Both (a) and (b)

Ans: (d)

Q 6: If two events A and B are independent, then

(a) A and the complement of B are independent

(b) B and the complement of A are independent

(c) Complements of A and B are independent

(d) All of these (a), (b) and (c)

Ans: (d)

Q 7: The probability of an event can assume any value between

(a) -1 and 1 (b) 0 and 1

(c) -1 and 0 (d) none of these

Ans: (b)

Q 8: If p:q are the odds in favour of an event, then the probability of that event is

(a) p/q (b) $\frac{p}{p+q}$

(c) $\frac{q}{p+q}$ (d) none of these

Ans: (b)

Q 9: For any two events A and B,

(a) $P(A)+P(B) \geq P(A \cap B)$ (b) $P(A)+P(B) < P(A \cap B)$

(c) $P(A)+P(B) > P(A \cap B)$ (d) $P(A) \times P(B) \leq P(A \cap B)$

Ans: (a)

Q 10: If two events A and B are independent, then $P(A \cap B)$ equals to

(a) P(A)+P(B) (b) P(A).P(B)

(c) P(A) . P(B/A) (d) P(B) . P(A/B)

Ans: (b)

Long Questions:

Q 1: There are two urns. The first urn contains 3 red and 5 white balls whereas the second urn contains 4 red and 6 white balls. A ball is taken at random from the first urn and is transferred to the second urn. Now another ball is selected at random from the second urn. Find the probability that the second ball is red.

Ans: 35/88

Q 2: A problem in probability was given to three students A, B and C whose chances of solving it are $1/3$, $1/5$ and $1/2$ respectively. What is the probability that the problem would be solved?

Ans: $11/15$

Q 3: A number is selected at random from the first 1000 natural numbers. What is the probability that the number so selected would be a multiple of 7 or 11?

Ans: 0.22

Q 4: The odds are 9:7 against a person A, who is now 35 years of age, living till he is 65 and 3:2 against a person B, now 45 years of age, living till he

is 75. Find the chance that atleast one of these persons will be alive 30 years hence.

Ans: $\frac{53}{80}$

Q 5: A bag contains 10 red and 8 black balls. Two balls are drawn at random. Find the probability that (a) both of them are red, (b) one is red and the other is black.

Ans: $\frac{5}{17}, \frac{80}{153}$



6.9 LET US SUM UP

- In this unit, we have discussed a few basic concepts of probability. We have defined the terms random experiment, trial, event and its different types, favourable cases, sample space etc.
- To discuss probability a few fundamental theorems exist, e.g., classical or mathematical or apriori definition of probability, addition theorem, multiplicative theorem etc.
- The Baye's theorem is also known as posteriori theorem. It was developed by Rev. Thomas Bayes.



6.10 FURTHER READING

1. Mukhopadhyay, P. (1996), *Mathematical statistics*. New Central Book Agency (P) Ltd.
2. Goon, A.M., Gupta, M.K. and Dasgupts, B. (1997), *An ouline of statistical Theory.* Vol. I, World Press.
3. Mukhopadhyay, P. (1991). *Theory of Probability.* New Central Book Agency (P) Ltd.
4. Gupta, S.C., Kapoor, V.K. (2009). *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons.

5. Feller, W. (1968). *An Introduction to Probability Theory and its Application*, Vol I, Wiley Eastern.



6.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No. 1: (i) True (ii) False (iii) True

Ans to Q No. 2: (i) P(A).P(B) (ii) $\frac{1}{2}$

(iii) independent (iv) 0.10



6.12 MODEL QUESTIONS

Q 1: A pair of unbiased dice is thrown. Find thje probility that (i) the sum of spots is either 5 or 10, (ii) either there is a doublet or a sum less than 6.

Q 2: In a group of 1000 persons, there are 650 who can speak Hindi, 400 can speak English and 150 can speak both Hindi and English. If a person is selected at random, what is the probability that he speaks (i) Hindi only, (ii) English only, (iii) only one of the two languages, (iv) atleast one of the two language?

Q 3: For two events A and B, let P(A) = 0.4, P(A+B) = 0.58 and P(B) = 0.3. Check whether A and B are independent or not.

Q 4: A can hit a target thrice in 5 shots, B twice in 5 shots and C thrice in 4 shots. They fire a volley. What is the probability that 2 shots hit?

Q 5: If A, B and C are mutually exclusive and exhaustive events, and P(A) = $\frac{1}{2}$. P(B) and P(B) = $\frac{2}{3}$. P(C), find P(A), P(B) and P(C).

*** ***** ***

UNIT 7: EXPECTATIONS AND MOMENTS

UNIT STRUCTURE

7.1 Learning Objectives
7.2 Introduction
7.3 Random Variable
7.3.1 Definition of Random Variable
7.3.2 Discrete Random Variable
7.3.3 Continuous Random Variable
7.4 Probability Distribution
7.4.1 Probability Mass Function
7.4.2 Probability Density Function
7.5 Mathematical Expectation and Variance
7.5.1 Mathematical Expectation
7.5.2 Properties of Mathematical Expectation
7.5.3 Variance of Random Variable
7.5.4 Properties of Variance
7.5.5 Illustrated Examples
7.6 Moments
7.7 Skewness
7.7.1 Karl Pearson's Co-efficient of Skewness
7.7.2 Boweley's Co-efficient of Skewness
7.8 Kurtosis
7.8.1 Measures of Kurtosis
7.9 Let Us Sum up
7.10 Further Reading
7.11 Answers to Check Your Progress
7.12 Model Questions

7.1 LEARNING OBJECTIVES

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

- distinguish between two types of experiments viz deterministic experiments and random experiments

- define what is a random variable and discuss when we get a discrete random variable or a continuous random variable
- have knowledge about probability distributions
- discuss the concept ot probability mass function and probability density function along with their properties
- know about mathematical expectation and variance of a random variable which are vital in the study of any distribution, along with their properties and applications
- distinguish between raw moments and central moments and how central moments can be expressed in terms of raw moments.
- realise how different measures of location and dispersion as discussed in earlier chapters can be expressed with the help of moments.
- discuss the concepts of skewness and kurtosis.

7.2 INTRODUCTION

The experiments that are performed can be classified into two broad categories. In one type of experiment the outcomes of the experiments remain same whenever and wherever the experiment is performed. In the second type of experiment, the outcomes are found to vary every time the experiment is performed.

The first types of experiments are known as deterministic experiments, whereas the second types are known as random experiments. As we have already discussed, in second type of experiment, the experimenter knows the set of all possible outcomes, but cannot say with certainity which outcome will turn out at a particular time. If we want to denote these various outcomes of a random experiment with the help of a variable, then it becomes a random variable.

In this unit we shall learn about random variables, the different types of random variables and their mathematical expectation and some related theorems. We shall also study about moments and try to understand the difference between raw moments, i.e. the moment about any arbitary point and central moment, the moment about mean.

7.3 RANDOM VARIABLE

Most of the random experiments we encounter have outcomes that can be interpreted in terms of real numbers, such as marks obtained by a student in an examination, number of accidents in a day etc. On the other hand, sometimes the outcomes of an experiment can take any value in an interval of real numbers. For example, the heights of students in a class can take any value within a specified range. Sometimes, the outcomes of a random experiment may be qualitative, e.g. tossing of a coin may result in head or in tail. We may denote head by '1' and tail by '0' and in this way the outcomes of a random experiment, quantitaive or qualitative can be expressed by a real number. these numerical values of the outcomes that changes from experiment to experiment are called random variables. So we can formally define a random variable as follows:

7.3.1 Definition of Random Variable

The real numbers which are associated with the outcomes of a random experiment are called random variables.

It can also be defined as a variable which takes a definite set of values with a probability associated with each of its values. That is why it is also called a chance variable or a stochastic variable.

Random variables may be of two types:

(i) Discrete random variables

(ii) Continuous random variables

Note: if X and Y are two random variables, then X+Y, X-Y are also random variables. Again, if, a and b are constants, then ax+y, x+by, ax+by are also random variables.

7.3.2 Discrete Random Variable

If a random variable x assumes only a finite number or countably infinite number of values, then it is called a discrete random variable. So a random variable x can take finite number of values $x_1, x_2, ..., x_n$ or it can take countably infinite number of values $x_1, x_2, ...$.

The number of letters received by a post office during a particular time period, the number of vehicles arriving at a toll bridge, the number of machines breaking down on a given day etc are some examples of discrete random variables.

7.3.3 Continuous Random Variable

If a random variable is such that it takes any value within a given interval, then it is called a continuous random variable. in this case we say that x takes any value in an interval, say. (a,b), i.e. $a \leq x \leq b$; i.e. x is a continuous random variable in the interval (a,b).

Amount of rainfall in a rainy season, heights of individuals, the time between arrival of customers at a service station etc are some examples of continuous random varible.

7.4 PROBABILITY DISTRIBUTION

The distribution obtained by taking all possible values of a random variable along with their respective probabilities is called a probability distribution. When the random variable is discrete we get discrete probability distribution and when the variable under consideration is continuous, we get a continuous probability distribution. The discrete probability distribution is called probability mass function and the continuous probability distribution is called probability density function.

7.4.1 Probability Mass Function

Let x be a discrete random variable which takes the values $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ with respective probabilities $p_1, p_2, \ldots\ldots\ldots, p_n$. Let, $p_i = P(x = x_i) = P(x_i)$, $i = 1, 2, \ldots\ldots\ldots$, n be the probability that x takes the value x_i. Then $p(x_i)$ is called the Probability Mass Function (p.m.f.) of x if it satisfies the following conditions.

(i) $p(x_i) \geq 0$ for all $i = 1, 2, \ldots\ldots\ldots\ldots$, n

(ii) $\sum_{i=1}^{n} p(x_i) = 1$.

In case, the discrete random variable x is countably infinite,

then, $p(x_i) = P(x = x_i)$, $i = 1, 2, ...$, is said to be the p.m.f. of x if

(i) $p(x_i) \geq 0$ for all $i = 1, 2, ...$

(ii) $\sum_{i=1}^{\infty} p(x_i) = 1$

Example 1: Let x denote the uppermost face while throwing a dice. Then x can take the values 1, 2, 3, 4, 5 or 6, each with probability 1/6. This can be written as,

$p(x_i) = P(x = x_i) = 1/6$, $i = 1, 2, ..., 6.$

Here, $p(x_i) > 0$

and $\sum_{i=1}^{6} p(x_i) = 1$

Thus $p(x_i)$ satisfies both the conditions for a p.m.f.

Example 2: Check whether the following probabilety distribution can be regarded as a Probabilety mass function and draw its graph,

x:	0	1	2	3	4
p(x):	0.10	0.20	0.40	0.20	0.10

Also find $P(x \geq 3)$

Solution : Here, $\sum_{x} p(x) = \sum_{x-0}^{4} p(x)$

= 0.10 + 0.20 + 0.40 + 0.20 + 0.10

= 1.

Also each of $p(x) \geq 0 \quad \forall x.$

The above distribution can be regarded as a probability mass functions.

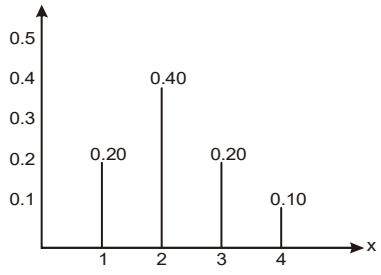


Fig. : probability mass function

Now $P(x \geq 3)$ =P(x=3) + P(x=4)

=p(3) + p(4)

= 0.20 + 0.10 = 0.30.

7.4.2 Probability Density Function

The probability distribution of a continuous random variable x is defined by $f(x)$ is called the probability density function (p.d.f.). Let, x be a continuous random variable in the interval (a,b), then the function f(x) = P(X= x) is said to be the p.d.f. of X, if it satisfies the following conditions.

(i) $f(x) \geq 0 \ \forall \ x$ in (a,b)

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

Example 3: Consider the function, $f(x) = \begin{cases} a, 0 \leq x \leq 5 \\ 0, otherwise \end{cases}$

For $f(x)$ to be a p.d.f., the condition $\int_a^b f(x)dx = 1$ must be satisfied, which is true if,

$$\int_0^5 a\,dx = 1$$

$$\Rightarrow \quad a = \frac{1}{5}$$

Since a>0, the function $f(x) \geq 0$. Thus f(x) satisfies both the conditions for a p.d.f.

Example 4: If $f(x) = kx$, $1 \leq x \leq 2$ is to be a probability density functino of the random variable X, find the value of k.

Solution: $\because f(x)$ is the p.d.f. of x, we must have,

$$f(x) \geq 0 \forall x$$

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

$$\Rightarrow k\int_1^2 x\,dx = 1$$

$$\Rightarrow k\left[\frac{x^2}{2}\right]^2 = 1$$

$$\Rightarrow k\left[2 - \tfrac{1}{2}\right] = 1$$

$$\Rightarrow k = \frac{2}{3}$$

7.5 MATHEMATICAL EXPECTATION AND VARIANCE

7.5.1 Mathematical Expectation

The mathematical expectation of a discrete random variable is the sum of the product of the values taken by the random variable and their respective probabilities. Let us consider the discrete random variable X, which assumes the values $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ with respective probabilities $p_1, p_2, \ldots\ldots\ldots, p_n$, then the mathematical expectation of the random variable X is denoted by E(x) and is given by

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

provided the series is convergent and $\sum_{i=1}^{n} p_i = 1$

If x is a continuous random variable with p.d.f. $f(x)$, $-\infty < x < \infty$, then the mathematical expectation of the random variable x is given by,

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx \text{, provided } \int_{-\infty}^{\infty} f(x) dx = 1$$

Note: The expectation of a random variable X represents the mean of the probability distribution of X.

Mathematical Expectation of a function of a random variable:

If $g(x)$ is a function of a random variable x, then the expectation of $g(x)$ is given by,

$$E[g(x)] = \sum_{i=1}^{n} g(x_i) p(x_i) = \sum_{i=1}^{n} g(x_i) p_i$$

where, x is a discrete random variable taking values $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ with respective probabilities $p_1, p_2, \ldots\ldots\ldots, p_n$.

Similarly, when x is a continuous random variable in the interval $(-\infty, \infty)$, with p.d.f. $f(x)$, $-\infty < x < \infty$, then,

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

7.5.2 Properties of Mathematical Expectation

Following are the important properties of mathematical expectation:

1. The expected value of a constant is the constant itself. i.e. E(c) = c, for every constant c.

2. The expected value of the product of a constant a and a random variable x is equal to a times the expected value of the random variable, i.e. E(ax) = a E(x).

3. The expected value of a linear function of a random variable is same as the linear function of its expectation, i.e. E (a+bx) = a+b E(x).

4. The expected value of the sum of the two random variables is equal to the sum of their individual expected values. i.e. E(x+y) = E(x) + E(y).

5. The expected value of the product of two independent random variables is equal to the product of their individual expected values. i.e. E(xy) = E(x) E(y).

7.5.3 Variance of a Random Variable

Variance is an important characteristic of a random variable. It provides the measure of dispersion of the random variable. The variance of a random variable x is denoted by V(x) or σ^2 and is given by,

$$\sigma^2 = v(x) = E\{X - E(X)\}^2 .$$

$$= E(X^2) - \{E(X)\}^2 .$$

Note: The positive square root of variance is the standard deviation.

7.5.4 Properties of Variance

1. The variance of a constant is zero. i.e. V(c) = 0.

2. The variance of the product of a constant and a random variable is equal to the variance multiplied by the square of the constant. i.e. $V(cx) = c^2 V(x)$.

3. The variance of the sum or difference of two independent random variables. i.e. $V(x \pm y) = V(x) + V(y)$.

7.5.5 Illustrated Examples

Example 5: A random variable x has the following probability distribution

x:	-2	-1	0	1	2	3
p(x):	0.1	0.1	0.2	0.2	0.3	0.1

Calculate E(X) and V(X)

Solution:

$E(X) = \sum_x xp(x)$

= -2 x 0.1 + (-1) x 0.1 + 0 x 0.2 +1 x 0.2 + 2 x 0.3 +3 x 0.1

= 0.8

$V(X) = E(X^2) - \{E(X)\}^2$

Now, $E(X^2) = \sum_x x^2 p(x)$

= $(-2)^2$ x 0.1 + $(-1)^2$ x 0.1 + 0 + 1^2 x 0.2 + 2^2 x 0.3 + 3^2 x 0.1

= 2.8

$\therefore V(x) = 2.8 - (-0.8)^2$

= 2.8 - 0.64

= 2.16

Example 6: The p.d.f. of a r.v. X is

$f(x) = \frac{2}{3}x, \ 1 \le x \le 2$,

find 9i) E(X), (ii) $P\left(\frac{5}{4} \le x \le \frac{7}{4}\right)$

Solution : We have, $f(x) = \frac{2}{3}x, \ 1 \le x \le 2$

(i) $\therefore E(x) = \int_a^b f(x)dx = \frac{2}{3}\int_1^2 x.x \ \ dx$

$= \frac{2^1}{3}\int_1^2 x^2 \ \ dx$

$= \frac{2}{3}\left[\frac{X^3}{3}\right]_1^2$

$$= \frac{2}{3}\left[\frac{8}{3} - \frac{1}{3}\right]$$

$$= \frac{14}{9}$$

(ii) $$P\left(\frac{5}{4} \le \times \le \frac{7}{4}\right)$$

$$= \int_{5/4}^{7/4} f(x) \quad dx$$

$$= \frac{2}{3}\int^{7/4} x \quad dx$$

$$= \frac{2}{3}\left[\frac{x^2}{2}\right]_{5/4}^{7/4}$$

$$= \frac{2}{3}\left[\frac{49}{32} - \frac{25}{32}\right]$$

$$= \frac{2}{3}.\frac{24}{32}$$

$$= \frac{1}{2}$$

Example 7: x is a random variable with p.d.f

$$f(x) = C(1-x), \qquad 0 < x < 1$$

fine (i) c (ii) E(3x+1)

Solution: Since f(x) is a p.d.f, we must have

$$f(x) \ge 0$$

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

Now, $$\int_a^b f(x)dx = c\int_0^1 (1-x)dx = 1$$

$$\Rightarrow C\left[x - \frac{x^2}{2}\right]_0^1 = 1$$

$$\Rightarrow C\left[1 - \frac{1}{2}\right] = 1$$

$$\Rightarrow C.\frac{1}{2}=1$$

$$\Rightarrow C=2$$

$$\therefore f(x)=2(1-x),\ 0 \le x \le 1$$

Now, E(X) $= \int_a^b x.f(x)dx$

$$= 2\int_0^1 x.(1-x)dx$$

$$= 2\int_0^1 x.(x-x^2)dx$$

$$= 2\left[\frac{x^2}{2}-\frac{23}{3}\right]_0^1$$

$$= 2\left[\frac{1}{2}-\frac{1}{3}\right]$$

$$= 2.\frac{1}{6}$$

$$= \frac{1}{3}$$

$\therefore$ (i) E(3x+1) $= 3E(X)+1$

$$= 3.\frac{1}{3}+1$$

$$= 2.$$

Example 8: In an office there are 10 HCL and 20 HP computers. One computer is selected at random. Find the probability distribution of selecting an HP computer. What is the expected value of selecting a HP computer.

Solution: Let X be a random variable such that,

x = 1, if a HP computer is selcted

= 0, if a HP computer is not selected.

$\therefore$ the probability distribution of x is,

x:	0	1
P(X=x):	10/30	20/30
	=1/3	= 2/3

$\therefore$ E (an HP computer is selected) = 0x1/3 + 1x2/3

= 2/3

Example 9: Eighty percent patients of a hospital are suffering from hypertension. One patient is randomly selected. Find the probability distribution of patients who are suffering from hypertension and find mean of the distribution.

Solution: Let x be a random variable such that,

x = 1, if the patient suffers from hypertension

= 0, otherwise

Hence the probability distribution of x is

x:	0	1
p(X=x):	0.2	0.8

and E(x) = 0 x 0.2 + 1 x 0.8

= 0.8

Example 10: The monthly demand for transistors have the following probability distribution

Demand (n):	1	2	3	4	5	6
Probability (p):	0.10	0.15	0.20	0.25	0.18	0.12

Determine the expected demand for transistors. If the cost of producing n transistors is given by C = 10,000+500n, determine the expected cost.

Solution: Expected demand for transistor

= E(n)

= $\sum np$

= 1 x 0.10 + 2 x 0.15 + 3 x 0.20 + 4 x 0.25 + 5 x 0.18 + 6 x 0.12

= 3.62

E (C) = E(10,000 + 500n) = 10,000 + 500 E(n)

= 10,000 + 500 x 3.62

= 11,810

Example 11: Proprieter of a food stall has invented a new item of food delicacy. He has calculated that the cost per piece will be Rs. 1, but he wants to sell it for Rs. 3 per piece. It is, however, perishable and the goods unsold at the end of the day is a total loss. If demand is a variable with the following frequency distribution, calculate his net profit if he makes 15 pieces.

No. of pieces demanded:	10	11	12	13	14	15
Probability:	0.07	0.10	0.23	0.38	0.12	0.10

Solution: Units produced = 15

Total cost price = 15 x 1 = Rs. 15

sell price per unit = Rs. 3

Unit demanded	Profit
10	Rs. (10x3) - 15 = rs. 15
11	Rs. (11x3) - 15 = Rs. 18
12	Rs. (12x3) - 15 = Rs. 21
13	Rs. (13x3) - 15 = Rs. 24
14	Rs. (14x3) - 15 = Rs. 27
15	Rs. (15x3) - 15 = Rs. 30

$\therefore$ Expected net profit = $\sum$ profit x probability

= 15x0.07 + 18x0.10 + 21x0.23 + 24x0.38 + 27x0.12 + 30x0.10

= Rs. 23.04



CHECK YOUR PROGRESS

Q 1: State whethre the following statemetns are true or false.

(i) The probability distribution of a random variable x is as follows

X = x:	1	2	3	
P(X= x):	1/4	1/2	3/4	(T/F)

(ii) If c is a constant then, (a) E(c) = c (b) V(c) = c (T/F)

(iii) If x and y are two independent random variables, then,

(a) V(x+y) = V(x) + V(y) (T/F)

(b) V(x-y) = V(x) - V(y) (T/F)

(iii) If x and y are two independent random variables, then,

E(xy) = E(x) E(y)

(iv) A continuous random variable can take any value within a given range.

Q 2: Define random variable. What are the important properties of a random variable.

.....

.....

.....

.....

7.6 MOMENTS

Moments are a special type of measure, which can reveal the characteristic of a distribution. They are very much important in the sense that, with the help of moments we can measure the central tendency of a set of observations, their variability, their asymmetry, and the height of the peak, their curve would make. Because of the great convenience in obtaining measures of the various characteristics of a frequency distribution, the calculation of the first four moments may well be made in the first step in the analysis of a frequency distribution. The letter 'μ' is used, generally, to denote the moments. Moments are called 'raw moments' when measured about any arbitary point 'a' and are called 'central moments' when measured about mean. We shall discuss them below.

Raw moments: The rth order raw moment about any point 'a' is,

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a)^r, \qquad N = \sum_{i=1}^{n} f_i$$

Putting r = 0, 1, 2, 3,..... in the above expression, we get

$$\mu'_0 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a)^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = 1$$

$$\mu'_1 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - a) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - a\frac{1}{N}\sum_{i=1}^{n} f_i$$

$$= \bar{x} - a = d \text{ (say)}$$

..... and so on

Central Moments: The rth central moment about mean $\bar{x}$ is called the central moment of order r, and is given by,

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x - \bar{x}\right)^r$$

Putting r = 0, 1, 2, 3,..... in the above expression we get,

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(x - \bar{x}\right)^0 = 1$$

μ_1 = [redacted] = [redacted]

= [redacted] - [redacted]

[redacted] = [redacted]

Now, we can express moments about mean in terms of moments about any point

We have, [redacted] = [redacted]

= [redacted]

= [redacted]

[redacted]

[redacted]

In particular,

[redacted]

[redacted]

[redacted]

We can also express skewness and kurtosis in terms of moments as follows:

Skewness =

Kurtosis =

Example 12:: The first three moments about the point x=10 are 5, 35 and 25 respectively. Final

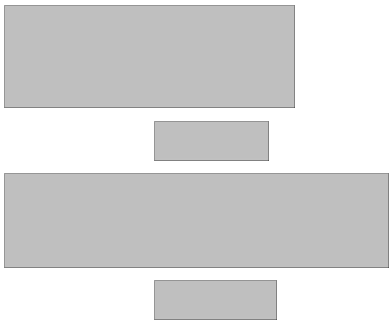
Solution: If

Now,

Example 13: The first four moment of a distribution about the value 4 of the variables are -1.5, 17, -30 and 108. Find the moments about mean and and

Solution: If usual rotation we have,

Moments about mean,



Note:

1) For a symmetrical distribution moments of odd order are zero.

2) For Normal distribution =0 and

3) First order raw moment about origin, i.e. is called the mean of the distribution.



CHECK YOUR PROGRESS

Q 3: fill in the blanks

(i) Moments are called _____ moments when measured about any arbitary point a.

(ii) Moments are known as _____ moments when measured about mean.

(iii) Second order central moment = _____.

(iv) For a symmetrical distribution all moments of odd order are _____.

(v) Skewness is measured in terms of moments as _____.

7.7 SKEWNESS

Literal meaning of skewness is lack of symmetry. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given frequency distribution. It helps us to determine the nature and extent of the concentration of the observations towards the higher or lower

values of the variable. In a symmetrical frequency distribution, which is unimodal, if the frequency curve is folded about the ordinate at the mean, the two halves so obtained will coincide with each other.

A frequency distribution, which is not symmetrical is called **asymmetrical** or **skewed**. In a skewed distribution, extreme values in a data set move towards one side or tail of the distribution, thereby lergthening that tail. A frequency distribution. For which the curve has a longer tail towards the right is said to be **positively skewed** (Fig 7.2) and if the longer tail lies towards the left, it is said to be negatively skewed (Fig 7.3).

Note : 1) For a symmetrical distribution, a Bell-shaped (normal curve) is obtained for which Mean = Median = Mode i.e. these three values fall at the same point (Fig 7.1)

Figure 7.1-7.3: Different Types of Skewness

Figure 7.1: Symmetrical Distribution

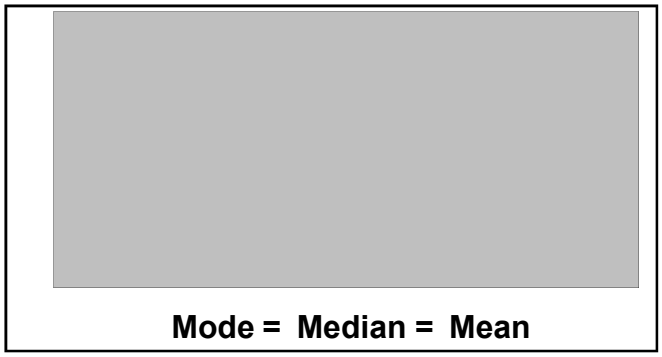


Figure 7.2: Positively Skewed Distribution

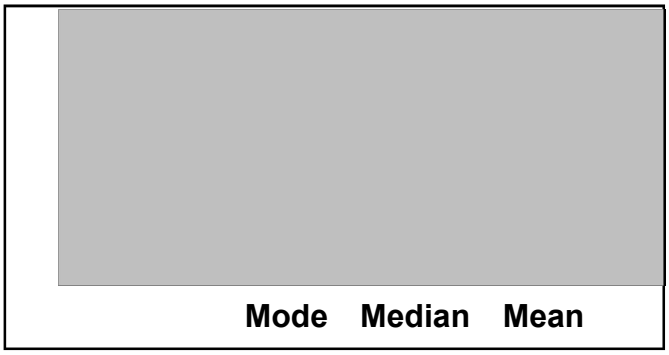
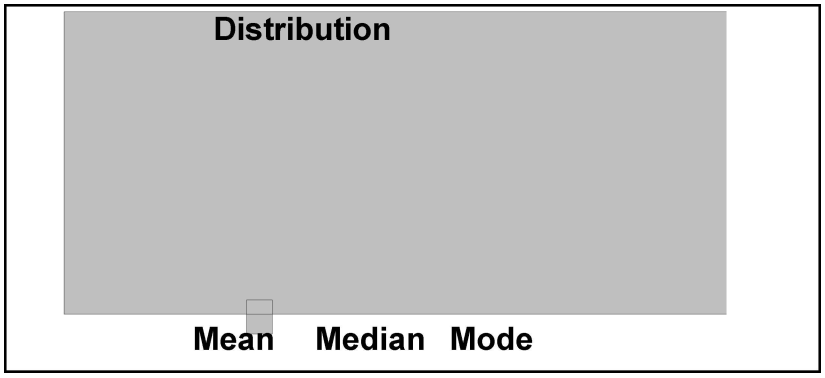


Figure 7.3: Negatively Skewed Distribution



2)For a positively skewed distribution A.M > Median > Mode, and for a negatively skewed distribution AM < Median < Mode.

3)For a skewed distribution, Quartiles Q_1 and Q_3 are not equidistant from the Median i.e.

$$Q_3 - \text{Median} \neq \text{Median} - Q_1$$

7.8 MEASURES OF SKEWNESS

For a skewed distribution, the difference between AM and Mode i.e (AM – Mode) can also be taken as a measure of skewness as for a positively skewed distribution AM > Mode and for a negatively skewed distribution AM < Mode. But the statistician observed that this measure may not be desirable as the diffence between mean and mode is expressed in the same unit as the distribution and therefore can not be used for comparing skewness of two or more distributions having different units of measurement. In order to overcome this shortcoming, relative measure of skewness (free from units of measurement), also commonly known as Co-efficient of skewness which is pure number suggested by karl pearson is preferred for measuring skewness.

7.7.1 Karl Pearson's Co-efficient of Skewness

Karl Pearson's Co-efficient of Skewness is given by the formula:



But quite often, Mode is ill-defined. In such a situation, we use the emperial relation between Mean, Mode and Median for moderately skewed distribution : Mean – Mode = 3 (Mean – Median)

In that case,



Remarks :

1. Theoretically, Karl Pearson's Co-efficient of Skewness lies between 

   moderately skewed distribution, SK_p lies between 

2. For a symmetrical distribution Mean = Mode = Median. Therefore, $SK_p = 0$. In other words, for a symmetrical distribution $SK_p = 0$.
3. SK > 0 if

For a positively skewed distribution, the value of Mean is the greatest of the three measures and value of Mode is the least.

On the other hand, SK < 0 if Mode > Median >

For a negatively skewed distribution the value of Mode is the greatest of the three measures and value of Mean is the leasl.

7.7.2 Boweley's Co-efficient of Skewness

There is another method suggested by Prof. Bowely known as **Bowleys'** Co-efficient of Skewness. It is given by



Remarks :

1. The values of Sk_3 lies between
2. This method of measuring skewness is quite useful when the distribution has open end classes.

Note : It is important to note that–

1. These two measures of Skewness namely Karl Pearson's and Bowley's co-efficient of Skewness can not be compared. On certain occasions it is possible that one of them gives a positive value while the other gives a negative value.
2. Out of the two measures, SK_p is more reliable due to the fact that sometime it may happen that $SK_B = 0$ but the distribution is not perfectly symmetrical.

Example 14 : Given for a distribution, pearson's measure of skewness is 0.4, AM = 30 and standard deviation = 8. Find Median and Mode.

Solution : Given, , $_p$= 0.4

We know that,





3.2 = 30 – Mode

Mode = 30 – 3.2 = 26.8

Also, 



3.2 = 90 – 3 Median

3 Median = 86.8

Median = 28.93

Example 15 : From the data given below, calculate the co-efficient of variation–

AM = 86, Median = 80, Pearsonian measure of skewness = 0.42.

Solution : Given $_p$ = 0.42.

We know that,







Co-efficient of variation = 



= 49.84%

Example 16: Given below are the Arithmetic mean, the Median and Standard Deviation of two distributions, Determine which distribution is more skewed.

(i) AM = 22, Median = 24, SD = 10

(ii) AM = 22, Median = 25, SD = 12

Solution : Here Pearson's Co-efficient of Skewness (SK_p) is applicable.



For distribution (i), 

For distribution (ii), 

For (i) $|SK_p| = 0.6$

For (ii) $|SK_p| = 0.75$

Since, $|SK_p|$ for (ii) > $|SK_p|$ for (i), hence distribution (ii) is more skewed.

Example 17: For a group 10 observations, sum of the observations is 452 and sum of squares of the observations is 24,270 and Mode is 43.7. Find the Pearson's co-efficient of Skewness.

Solution : Given n = 10, 

Mode = 43.7

Mean 







Example 18: In a frequency distribution, the co-efficient of Skewness lased on quartiles is 0.6. If the sum of upper and lower quaitiles is 100 and Median is 38, find Q_3.

Solution : Given $SK_B = 0.6$

$Q_3 + Q_1 = 100$

Median = 38

We know that, 





Given $Q_3 + Q_1 = 100$

$(Q_3 - Q_1) + (Q_3 + Q_1) = 40 + 100$

$2Q_3 = 140$

$Q_3 = 70$

Example 19: Calculate Karl Pearson's Co-efficient of Skewness from the data given below.

Wt (Kg) :	10	20	30	40	50	60	70
Frequency:	1	5	12	22	17	9	4

Solution : Table for calculation of co-efficient of Skewness

Wt (kg)	Frequency	d = x –A	$d'=d/10$	fd'	fd'^2
(x)	(f)				
10	1	–30	–3	–3	9
20	5	–20	–2	–10	20
30	12	–10	–1	–12	12
40	22	0	0	0	0
50	17	10	1	17	17
60	9	20	2	18	36
70	4	30	3	12	36
	70 = N				

Let assumed mean = A = 40

Here i = 10





= 43.14 kg.

Mode = 40 (By inspection)





Example 20: From the data given below, calculate Karl Pearson's co-efficient of Skewness.

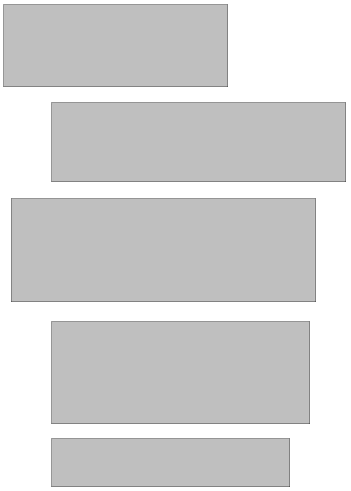
Solution : Table for calculating Karl Pearson's co-efficient of Skewness.

Ht (cm) : 130-134 135-139 140-144 145-149 150-154 155-159 160-164

f : 3 12 21 28 19 12 5

Ht (cm)	(f)	Mid value(x)	d = x–A	$d'=d/5$	fd'	fd'^2	Class Boundarie
130-134	3	132	–15	–3	–9	27	129.5–134.5
135–139	12	137	–10	–2	–24	48	134.5–139.5
140–144	21	142	–5	–1	–21	21	139.5–144.5
145–149	28	14.7	0	0	0	0	144.5–154.5
150–154	19	152	5	1	19	19	149.5–154.5
155–159	12	157	10	2	24	48	154.5–159.5
160–164	5	162	15	3	15	45	159.5–164.5
	N= 100					$\sum fd'^2=208$	

Let A = 147 Here i = 5



= 7.21 cm

Here Modal Class is 144.5 – 149.5

Mode =







= 146.69 cm



Example 21 : Fine the most suitable measure of Skewness from the data given below.

Age (yrs) : Below 20 20–25 25–30 30–35 35–40 40–45 45–50

No. of

employee 13 29 46 60 112 94 45

50 and above 21

Solution : This is an open-ent series. So. the most suitable measure of Skewness is Bowley's Co-efficient of Skewness.

Age (yrs)	No. employee (f)	c.f.
Below 20	13	13
20 – 25	29	42
25 – 30	46	88
30 – 35	60	148
35 – 40	112	260
40 – 45	94	354
45 – 50	45	399
50 and above	21	420
	N = 420	

We know that, Bowley's coefficiant of Skewness



Here N = 420

Q_1 = Value of

Median = Value of

Q_3 = Value of

Q_1 class is 30 – 35

Median class is 35 – 40

Q_3 class is 40 – 45

= 30 + 1.42 = 31.42 yrs

Median

= 34 + 2.77 = 37.77 yrs

and

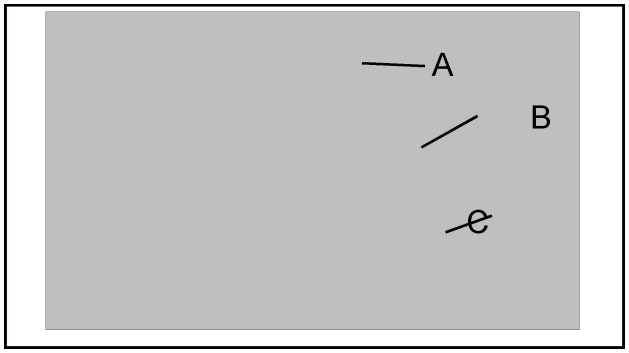
= 42 + 2.93 = 42.93 yrs

7.8 KURTOSIS

So far we have studied three measure viz central tendency, dispersion and skewness to describe the characteristics of a frequency distribution. However, even if we know all these three measures, we all not

in a position to charcterise a distribution completely. The diagram given below will clarify the situation.

Figure 7.4: Lapto kurtic, meso kurtic and platy kurtic curves



A – Lapto Kurtic
B – Meso Kurtic
C – Platy Kurtic

Mean

All these three curves are symmetrical about the mean and have same variation (range). But the shape and nature of the middle part of the curve are different i.e the peakedness of the three curves are not same. So. to identify a distribution completaly, we need one more measure which is called **'Kurtosis'**.

The word **Kurtosis** comes from a Greek word meaning 'humped'. In statistics, **Kurtosis refers to the degree of flatness or peakedness of the frequency curve.**

Curve of type B which is neither flat peaked is known as normal curve and shape of its hump is accepted as a standard one. This curve is termed as **mesokurtic.** The curves of type A, which is more peaked than the normal curve are known as **leptokurtic** and the curves of type C, which are flatter than the normal curve are called **platykurtic**

7.8.1 Measures of Kurtosis

As a measure of Kurtosis, Karl Pearson gave the coefficient Beta two ( ) or its derivative Gramma two (

For a normal curve (mesakurtic curve), $\beta_2 = 0$. For a lepto kurtic curve, $\beta_2 > 0$ and for a platy kurtic curve, $\beta_2 < 0$ or $r_2 < 0$.

Note : A British statistician W. S. Gosset explained the use of the terms platykurtic and leptokurtic as : 'platykurtic curves like platyplus, are squat with short tails; leptokurtic curves are high with long tails like the kangaroos noted for leaping.

Example 22 : The first four moments of a distribution about the value 5 and 2, 20, 40 and 50. Do you think that the distribution is platykurtic?

Solution : Here A = 5,

Mean

Now,

Variance

$= 40 - 3 \times 2 \times 20 + 2(2)^3 = -64..7$

$= 50 - 4 \times 2 \times 40 + 6 \times 20 \times (2)^2 - 3(2)^4 = 162$

Since

Example 23 : The first four moments of a distribution about the origin are 1, 4, 10 and 46 respectively. Calculate the central moments and indicates the nature of the distribution.

Solution : given A = 0,

Mean

Variance

$= 10 - 3 \times 4 \times 1 + 2(1)^3$

$= 0$

$= 46 - 4 \times 10 \times 1 + 6 \times 4 \times 1 - 3 = 27$

Karl Pearson's moment co-efficient of Skewness is given by

Since,

Again, Karl Pearson's measure of Kurtosis is given by

and

Since,

Mean

and

Example 24 : If

and

Solution: given

Variance

Now,

Also

and

Hence the given distribution is leptokurtic.



CHECK YOUR PROGRESS

Q 4: Explain meaning of skewness using sketches of frequency curves. How does skewness different from dispersion.

Q 5: Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and why?

Q 6: How is Kurtosis measured ?

.....

.....

.....

.....



7.9 LET US SUM UP

- The experiments can be classified into two broad categories - deterministic experiment and random experiment.
- The real numbers which are associated with the outcomes of a random experiment are called random variables.
- Random variables may be of two types - discrete random variable and continuous random variable.
- The distribution obtained by taking all possible values of a random variable along with their respective probabilities is called a probability distribution.
- The probability distribution for discrete random variable is called probability mass function (p.m.f.) and the same for continuous random variable is known as probability density function (p.d.f.).
- The mathematical expectation of a discrete random variable is the sum of the product of the values taken by the random variable and their respective probabilities.
- The expectation of the random variable x represents the mean of the probability distribution of x.
- The variance of a random variable provides a measure of dispersion of the random variable.
- The expectation of a constant quantity is the constant itself, whereas variance of a constant is zero.
- The expectation of the sum of two or more random variables is the sum of the individual random variables.

- The expectation of the product of two independent random variables is the product of their expectations.
- The variance of the addition or subtraction iof two independent random variables is equal to the sum of their variances.
- Moments are a special type of measure which can measure the characteristic of the distribution - like the central tendency of a set of observations, their variability, symmetry, peakedness etc.
- Moments are of two types - Raw moments and Central momnets.
- Moments are termed as raw moments when measured about any arbitary point, say, a and central moments are those moments when measured about the mean of the distribution.
- Central moments can be expressed in terms of raw moments and vice versa.

7.10 FURTHER READING

1. Mukhopadhyay, P. (1996), *Mathematical statistics*. New Central Book Agency (P) Ltd.
2. Goon, A.M., Gupta, M.K. and Dasgupts, B. (1997), *An ouline of statistical Theory*. Vol. I, World Press.
3. Mukhopadhyay, P. (1991). *Theory of Probability*. New Central Book Agency (P) Ltd.
4. Gupta, S.C., Kapoor, V.K. (2009). *Fundamentals of Mathematical Statistics*, Sultan Chand & Sons.
5. Feller, W. (1968). *An Introduction to Probability Theory and its Application*, Vol I, Wiley Eastern.

7.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1:

(i) False (Total probability exceeds 1)

(ii) (a) True, (b) False

(iii) (a) True (b) False

(iv) True

(v) True

Ans to Q No 2:

The real numbers which are associated with the outcomes of a random experiment are called random variables.

The properties of random variable are as follows:

(i) If X is a random variable then, 1/x, x^2 are also random variables.

(ii) If x is a random variable and a and b are constants then ax+b is also a random variable.

(iii) If x and y are two random variables, then x+y and x-y are also random variables.

Ans to Q No 3:

(i) raw (ii) central (iii) variance

(iv) zero (v)

Ans to Q No 4: (i) Equal (ii) Laptakurtic (iii) symmetrical. (iv) negatively

(v) peakedness/flatness. (vi) (vii) $_2$.

Ans to Q No 5: (i) True (ii) False (iii) False (iv) True

(v) True (vi) False.

Ans to Q No 6: (i) True (ii) False

7.12 MODEL QUESTIONS

Q 1: Fill in the blanks

(i) Heights or weights of individuals can be represented by a _____ random variable.

(ii) If x and y are two random variables then ax+by is a _____, where a and b are constants.

(iii) If x and y are independent, then E(xy) = _____.

(iv) V(x y) = V(x) + V(y), when x and y are _____ variables.

(v) V(ax+b) = _____, where a and b are constants.

(vi) E(a) = _____ and V(a) = _____, where a is a constant.

(vii) Skewness is given by

(viii)

(ix)

Q 2: Can you find the mathematical expectation of a random variable x, if x takes the values 1,2,3 with corresponding probabilities 1/4, 1/2 and 3/4.

Q 3: Discuss discrete and continous random variables with examples.

Q 4: What do you mean by mathematical expectation of a function of a random variable. State addition and multiplication law of mathematical expectation.

Q 5: A random variable x has the following probability distribution,

x:	-2	-1	0	1	2	3
f(x):	0.1	2k	0.2	2k	0.3	2k

Find the value of k.

Q 6: Write some important properties of mathematical expectation.

Q 7: Let x be a random variable with density function

f(x) = c(1-x), 0<x<1.

Q 8: What are moments? Distinguish between raw moments and central moments.

Q 9: Calculate Karl Peason's Co-efficients of Skewness for the given distribution.

Marks less than:	20	40	60	80	100
No. of students:	18	40	70	90	100

Q 10: Calculate Bowley's Co-efficient of Skewness.

Marks :	0–10	10–20	20–30	30–40	40–50
No. of students:	5	9	12	8	6

Q 11: The arithmetic mean of a certain distribution is 5. The second and the third moments about the mean are 20 and 140 respectively. Find the third moment of the distribution about zero.

*** ***** ***

UNIT 8: THEORETICAL PROBABILITY DISTRIBUTIONS (DISCRETE VARIABLE)

UNIT STRUCTURE

8.1 Learning Objectives

8.2 Introduction

8.3 Moment Generating Function

- 8.3.1 Definition
- 8.3.2 Generation of Moments
- 8.3.3 Properties of Moment Generating Function
- 8.3.4 Illustrated Examples

8.4 Binomial Distribution

- 8.4.1 Derivation of Binomial Distribution
- 8.4.2 Definition
- 8.4.3 Moments of Binomial Distribution
- 8.4.4 Moment Generating Function of Binomial Distribution
- 8.4.5 Fitting of Binomial Distribution
- 8.4.6 Properties of Binomial Distribution
- 8.4.7 Illustrated Examples

8.5 Poisson Distribution

- 8.5.1 Poisson Distribution as a Limiting Case of Binomial Distribution
- 8.5.2 Moments of Poisson Distribution
- 8.5.3 Moment Generating Function of Poisson Distribution
- 8.5.4 Fitting of Poisson Distribution
- 8.5.5 Properties of Poisson Distribution
- 8.5.6 Application of Poisson Distribution
- 8.5.7 Illustrated Examples

8.6 Let Us Sum Up

8.7 Further Reading

8.8 Answers To Check Your Progress

8.9 Model Questions

8.1 LEARNING OBJECTIVES

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

- obtain moment generating function for any given distribution
- generate moments from any probability distribution
- learn some important properties of moment generating function
- derive binomial distribution from a coin tossing experiment
- derive poisson distribution as a limiting form of binomial distribution
- compute moments of binomial and poisson distribution
- fit binomial and poisson distribution to real life data
- describe the application of the concept of these probability distributions to real life data.

8.2 INTRODUCTION

So far we have brought out the nature of discrete and continuous probability distributions in general and also have defined their mean and variances. In this present unit we shall discuss about how the moments of a particular distribution as discussed in chapter 6, can be generated by a special kind of function called the moment generating function.

Again we know that the rule generating the probability distribution is the same whether we want to find the probability of the number of heads when three coins are lossed simultaueously or for the number of red balls in case of an urn containing 3 red and 2 black balls, out of which, say, two balls are drawn at random with replacement. Given this basic characteristic of a discrete distribution, our main concern is to examine the outcoms generated by an experiment and choose a distribution that best describes those outcomes.

Under this consideration we shall discuss some discrete probability distributions applicable to many real life situations in the subsequent sections of this unit. The distributions we are going to discuss here are binomial and poisson distributions.

8.3 MOMENT GENERATING FUNCTION

The moments of a probability distribution play an important role in theoretical as well as in applied statistics. With the help of moments we can measure the central tendency of a set of observations, their variability, their asymmetry and the height of the peak their curves would make. That is the knowledge of moments enable us to determine the probability distribution completely. So it would be very useful if a function could be found which will help in generating moments. One such function is moment generating function. So we can define a moment generating function as follows :

8.3.1 Definition

A function that is used to generate moments of a random variable X is called the moment generating function, denoted by Mx (t) and is given as

.

When X is a discrete random variable with p.m. f. p(x)



When x is a continous random variable in the interval (a,b) with p.d.f. f(x),

.

8.3.2 Generation of Moments

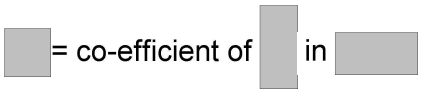
We have, 



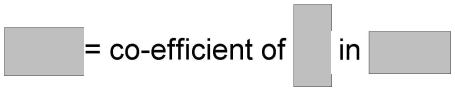


 = co-efficient of t in 

 = co-efficient of  in 



Thus in general we see that,



Another convenient way of obtaining moments is to differentiate

So if we differentiate

then put t = 0, we get

i.e.

Differenting

. w

So in general, we get the raw moment of order r,



8.3.3 Preperties of Moment Generating Function

(i) If X is a random variable and c is a constant, there,



Proof : L. H. S =

= R.H.S.

(ii) Additive property of moment generating function.

The moment generating function of sum of independent random variables is equal to the product of their moment generating functions. In symbol, if X_1, X_2, ..., X_n are independent random variables then,



Proof : L.H.S. = 






are independent

= 

= R.H.S.

(iii) Effect of change of origion and scale on m. g. f.



Proof : 





= 

= R.H.S.

Note : (i) The moment generating function does not always exist.

(ii) In some cases the moment generating function exists, but it cannot generate moments.

8.3.4 Illustrated Examples

Example 1: Let X be a random variable with p. m. f.

, 

find the moment generating function and hance find its mean and variance.

Solution :









*

Mean = 





















*The expansion used here is called binomial expansion with negative index, which is 

 variance 









Example 2: Final moment generating function of a continuous random variable x having the p.d.f.



Hence find its mean and variance.

Solution : 







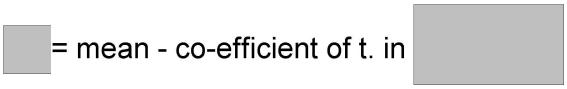









= mean - co-efficient of t. in


= mean - co-efficient of t. in


variance



CHECK YOUR PROGRESS

Q 1: State true or false.

(i) Since M_x (t) generates moments, it is called moment generating function. (T / F)

(ii) If X an Y are independent random variables then [redacted] (T / F)

(iii) Moment generating function always exits. (T / F)

(iv) A moment generating function, if exists, determences the distribution function. (T / F).

8.4 BINOMIAL DISTRIBUTION

The binomial distribution, as derived by James Bernoulli in the year 1700, deals with population whose elements or outcomes can be devided into two catigories with reference to presence or absence of a particular attribute or characteristic. For example, the answer to a question may be correct or incorrect, results in an examination may be pass or fail, the factory workers may be educated or uneducated.

That is the outcome in each of these situations may be viewed in terms of presence or absence of an attribute. That is we consider only the two outcomes one possessing and the other not possissing the attribute. A convenient method is to denote them as ‘success’ and ‘failure’ according as the presence or absence of the attribute.

The mathematical background for binomial distribution describes descrete date resulting from an experiment called. Bernoulli trials, which are repeted finite number of times. Let us now device binomial distribution from a finite number of Bernoulli trials.

Note : A random experiment whoss outcomes has been classified into two categories, called ‘success’ and ‘failure’ is called Bernoulli trial. For example, tossing of a coin may result in a head or in a tail etc.

8.4.1 Derivation of Binomial Distribution

Consider a coin tossing experiment which is repeated, say, n number of times. When we get a head we call it a success and when we get a tail, we call it a failure, success is denoted by 's' and failure by 'f'. Probability of success is p and that of failure is q.

P(s) = p, P(f) = q, q = 1 – p.

suppose we are interested in getting x success out of n trials. Now the probability of x successes and consequently (n-x) failures in n independent trials, in a specified order, say, first x trials succes and the remaining (n-x) trials failures

P(ss . . . s f f . . . f) = P(s) P(s) . . . P(s) P(f) . . . P(f)

the trials are independent

=

But this is simply one way of getting the probabilety of x successes out of n trials. As we know, x successes in n trials can occur in n_{C_X} ways, and the probability for each of these cases is $p^x q^{n-x}$, Hence by the addition theorem of probability the probability of getting x success in n trials is given by.

times

Which is the probability mass function of binomial distribution. Again, , the number of successes out of n trials can take values 0, 1, 2, ..., n. Therefore, we may define a binomial distribution as follows.

8.4.2 Definition

A random variable X is said to follow binomial distribution if it assumes non-zero values and if its p.m.f. is given by,

0, otherwise.

The two constante n an p are the parameters of this dist^{n}. It is also abbreviated as B(n, p).

Note : The probability of getting 0 success. 1 success, ..., n successes in this case, i.e. successive terms in the expansion of , a binomial expansion, that is why this distribution is also termed as 'binomial distribution'.

8.4.3 Moments of Binomial Distribution

We are interested mainly in the first two moments of benomial distribution which are obtained as follows.

=

=

Hence the mean of binomial distribution is n p.

Again,

.





= variance

=

=

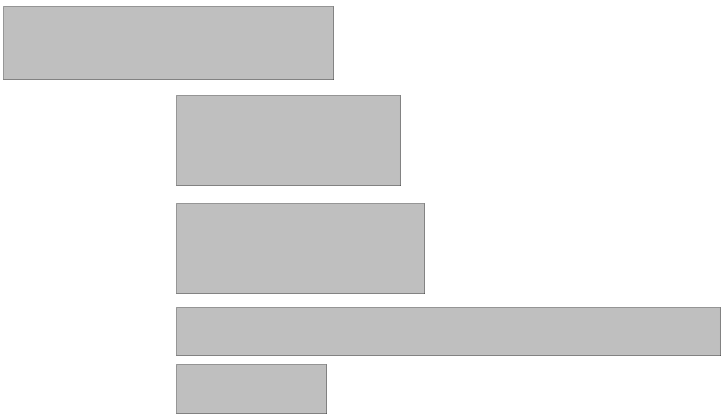
=

So the variance of binomial distribution is

Remark : For binomial distribution mean is always greater than variance. This is an important characteristic of binomial distribution.

8.4.4 Moment Generating Function of Binomial Distribution

If X ~ B (n, p), then



which is the moment generating function of binomial dist^{n}.

8.4.5 Fitting of Binomial Distribution

The recurrence relation for probabilities of binomial distribution is,

Putting we get

to calculate only

For fitting a binomial distribution, we need to find the values of and

. is generally known, and

where np is the mean of the given set of data.

After calculating the value of p, we calculate the probabilities as discussed above. Now multiplying these probabilities by N, the total frequency, we get the expected frequencies for distribution with the help of an example.

Note : Calculation of expected frequency is called fitting.

Example 3: 3 coins are tossed 64 times and the following distribution of number of heads were obtained.

Number of heads	:	0	1	2	3
Frequency	:	10	18	20	16

Fit a binomial distribution under the hypothesis that the coins are unbiased.

Solution : Since the coins are unbiased

[in case coins are biased be calculated from mean]

Here, N = 64,

Now,

Now using recurrence relation,

we calculate the probabilities as,

Hence the expected frequenties are,

	Expected frequency
0	

1

2

3

8.4.6 Properties of Binomial Distribution

(i) The binomial distribution is a discrate distribution where the random variable X takes the values 0, 1, 2, ..., n.

(ii) The binomial distribution has two parameters, and

(iii) The mean is equal to np and variance is given by

(iv) The mean of binomial distribution is always greater than variance.

(v) If x_1 and x_2 follows binomial distribution with parameters and $_1+ x_2$ follows binomial distribution with parameter additive property of binomial distribution.

(vi) Binomial distribution with parameters distribution if is very largs, is very small and

(vii) Binomial distributions also tends to Normal distribution if very large and neither nor

8.4.7 Illustrated Examples

Example 4: A r. v. (r.v. : random variable) X follows binomial distribution with mean 5/3 and P(x = 2) = P (x = 1). Find variance, P(x = at least 1), P(x = at most 1).

Solution : If X follows binomial distribution with parameters n an p, the p.m.f. of X is,

, = 0, 1, 2, . . . , n;

Given, mean = =

and P(x = 2) = P(x = 1)

Again,

Hence, variance

P(x = at least 1) =

P(x = at most 1) = P(x=0) + p(x=0)

Example 5: A box contains 3 red balls and 2 white balls. In a sample of 3 balls drawn, find the probability of getting (i) one red ball, (ii) two red balls, if the balls are drawn with replacement.

Solution : p = Prob of drawing a red ball

prob. of getting x red balls in a sample of 3 balls is,

,

Now, (i)

= p(1)

=

(ii) p(2) =

Example 6: Ten coins are thrown simultaneously, Find the probability of getting at least seven heads.

Solution: p = prob. of getting a head =

q = prob. of not gettig a head =

probability of getting x heads in a throw of 10 coins is

prob. of getting atleast 7 heads is given by,

Example 7: The incidence of occupational disease in an industry is such that the workmen have a 20% chance of suffereing from it. What is the probability that out of six workmen, 4 or more will contract the disease?

Solution: prob that the workmen will suffer from

disease

n = 6

using the probability law we get,

Example 8: A multiple choice test consists of 10 questions with 3 answers to each questions, of which only one is correct. A student answers answers each question by rolling a dice and checking the first answer if he gets 1 or 2, second answer if the gets 3 or 4 and the third answer if he gets 5 or 6. To get a distinction, the student must secure at least 80% correct answer. If there is no negative marking, what is the probability that the student secures a distinction.

Solution: Probability of getting correct answer to a question is,

and

Required probability of getting a distinction (i.e getting correct answers of 8 questions, out of 10) is,

6. Probability of failure in statistics examination is 15%, 20 batchs of 5 students each appear for the examination. Obtain in how many batches 3 or more students would pass.

Sol^{n} : Given q = probability of failure = 15% = 0.15

p = probability of success = 0.85

n = 5.

If X denote the number of students who would pass, then

, x = 0, 1, . . ., 5.

We need

$= 10x\ (.85)^3\ (.15)^2 + 5(0.5)^4\ (0.15) + (0.15)^5$

$= 0.1381 + 0.3915 + 0.4437$

$= 0.9733.$

Required number of batches in which 30x more students would pass $= 20 \times 0.9733$

$= 19.468$



CHECK YOUR PROGRESS

Q 2: If what is E(X)?

Q 3: Examine the following statement and add your comment, "The mean of a binomial distribution is 5 and s.d. is 3".

Q 4: What are the parameters of binomial distribution?

Q 5: Find the varience of the binomial distribution

, = 0, 1, 2, . . ., 10.

Q 6: Define Bernoulli treats. Find its distributions and also its mean and variance.

Q 7: Derive Binomial distribution

or

Obtain the probability of exactly successes in independent Berioulli treats.

8.5 POISSON DISTRIBUTION

The Poisson process was first development by French mathematician and physicist serieon Devis Poisson, who published it in the year 1837. In the class of univariate distributions, the poisson distribution is perhaps the most useful one, because of its wids applicability in the description of natural phenomena, for example, demand for service in a big departmental store, major accident rate per day in a certain highway etc. A typical application of poisson distribution is for analysing queing (or waiting live) problems in which arriving customers during an interval of time arrive independently and the number of arrivals depends on the length of the time interval.

8.5.1 Poisson Distribution as a Limiting Case of Binomial Distribution

The binomial probability law of rare events gives rise to poisson probability law. That is poisson distribution can be derived as a limiting case of binomial distribution under the following conditions.

(i) n, the number of independent Bernoulli trials is very large, i, e

(ii) p, the probability of success in each trial is very small, i.e.

(iii) np = (say), is finite, where

the above conditions, we can derive the p.m.f. of poisson distribution as follows:

Let us consider the p.m.f. of poisson distribution as,

, x = 0, 1, 2, . . ., n; q = 1 – p.

Now,



Which is the p.m.f of poisson distribution. We give below the formal definition of poisson distribution as,

Definition :

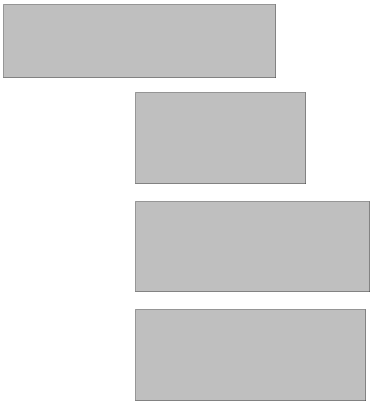
A discrete random variable X is said to follow poisson distribution, if it assumes non -ve values only and its p.m.f is given by,

 , x = 0, 1, 2, . . .,

= 0,

8.5.2 Moments of Poisson Distribution

The arithmatic mean of poisson distribution is given by,







the mean of poisson distribution is



Now,




The exponential exponsion,









Now,



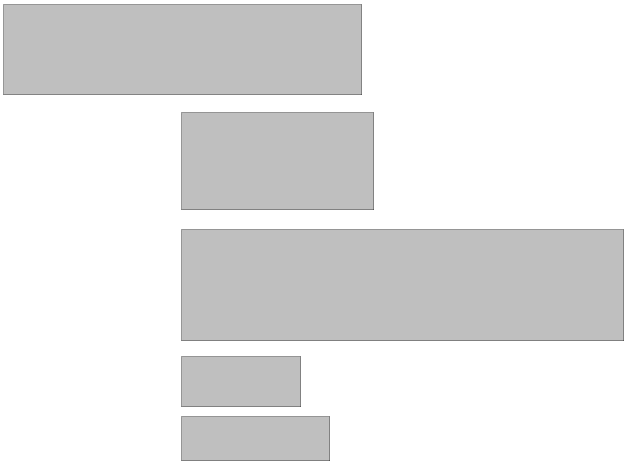


Hence the variance of poisson distribution is also

Remark : Poisson distribution is the only discrate distribution where mean is equal to variance.

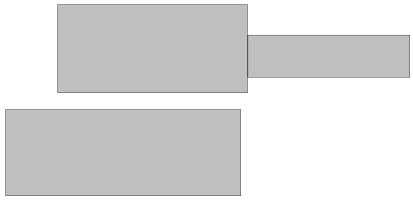
8.5.3 Moment Generating Function of Poisson Distribution

The moment generating function of poisson distribution is given by,

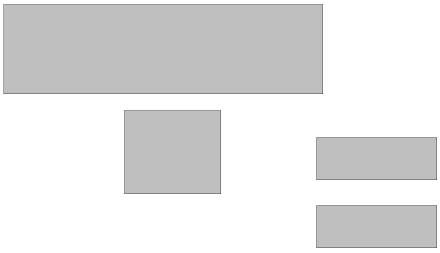


8.5.4 Fitting of Poisson Distribution

The p.m.f. of Poisson distribution with parameter is,



Now,



Hence, the recurrence.....

The recurrence relation for probabilities of poisson distribution is,

 = 0, 1, 2, ... (1)

Putting x = 0, 1, 2, ... , in (1) we get the probabilities p(1), p(2), ... , etc. so we need to calculate p(0) only, which is,

.

where,

data to be fitted.

Ultimately, the expected frequencies are calculated by

multiplying the corresponding probabilities by N, the total frequency.

We shall explain the fitting of poisson distribution to a given data set.

Example 9: 200 drivers made the following road accidents

Accidents :	0	1	2	3	4
No. of drivers :	122	60	15	2	1

Fit a theoretical distribution to the data.

Solution : The suitable theoretical distribution to be fitted here may be the poission distribution,

Let, = number of road accidents

= no. of drivers.

Total number of drivers = N = 200

Here,

Using recurrence relation we get,

the expected frequencies are,

= 200 x 0.60653 = 121.306

= 200 x 0.30327 = 60.654

= 200 x 0.07582 = 15.164

= 200 x 0.01264 = 2.528

= 200 x 0.00158 = 0.316

Total frequency = 200

8.5.5 Properties of Poisson Distribution

(i) Poisson distribution is a discrete distribution where the random variable X takes the values 0, 1, 2, ... , .

(ii) The poisson distribution has only one parameter

(iii) The mean and the variance of poisson distribution are equal to .

(iv) The poisson distribution has only one or two modes.

(v) Poisson distribution can be obtained as a limiting case of binomial distribution.

(vi) If X and Y are independent poisson variates with paremeters and
parameter

8.5.6 Application of Poisson Distribution

Physical situations in daily life where poisson distribution is applicable are,

(i) Number of telephone calls recieved per minute at a particular switch board.

(ii) Number of accidents in a city per unit time.

(iii) Number of defective parts produced in a factory per unit time.

(iv) Number of printing mistakes in each page of a book.

8.5.7 Illustrated Examples

Example 10 : Find the mean and variance of poisson distributionsuch that, P(x = 2) = P(x = 1)

Solution : It X follows poisson distribution with parameter

then, , = 0, 1, 2, ...

Now, p (x = 2) = p (x = 1)



the mean and variance of poisson distribution is equal to

Example 11 : The number of ticketless travallers detected per checking follows poisson distribution with mean 10. Find the probability of getting two ticketless travellers.

Solution : Let X denote the no. of ticketless travellers in a checking. so X follows poisson distribution with given mean.

, = 0, 1, 2, ...

P(getting two ticketless travellers during a checking)

= P(x = 2)

= 0.0025.

Example 12: What probability model is appropriats to describs a situation where 100 misprints are distributed randomly throughout the 100 pages of a book? For this model what is the probability that a page observed at random will contain at least 3 misprints.

Solution : Since 100 misprints are distributed randomly throughout the 100 pages of the book, the probability that there being a misprint is

n = 100 pages is quite large.

Hence poisson distribution is best fitted in this case.

Here,

Probability that there are alteast three misprints in a page is











Example 13: A factory producting machine parts finds that on an average there are 5 defective items in a lot of 1000 units. The units are packed, 500 in packet. Calculate the number of packets containing 3 defective and 4 defectives in a consignment of 2000.

Solution: Here n = 500, is large

p = 

N = 2,000.

Hence, the distribution can be taken as Poisson distribution.

Here, 

Again, the number of packets with x defectives is

given by, 

No. of packets with 3 defectives



=

No. of packets with 4 defectives



Example 14: A car fire firm has two cars which it hired out day by day. The number of demands for a car on each day is distributed as Poisson variate with mean 1.5. Calculate how many days in a month (i) neither car is used and (ii) some demand is refused.

Solution: Probability of days on which there are x demands for a car = p(x)

= 

since the number of demands for a car on any day is a Poisson variate with mean 1.5.

(i) the number of days in a month, on which neither car is used is given by,

(ii) No. of days in a month, when some demand is refused is given by,



CHECK YOUR PROGRESS

Q 8: Fill in the blanks :

(i) For a poisson distribution mean is ..... to variance (less/equal/greater)

(ii) The s.d. of a poisson distribution with mean 4 is ..... (4/2).

(iii) For a poisson distribution , x = 0, 1, 2, ... the variance is (1.5/2.5).

(iv) If x follows poisson distribution then its m.g.f is

(v) If are poisson variates then the distribution of is also a ..... variate. (Binomial / Poisson)

Q 9: "Poisson distribution is also known as probability distribution of rare events" – Explain.

Q 10: Give some examples of poisson distribution.



8.6 LET US SUM UP

- A function, which is used to generate moments of a probability distribution is known as moment generating function.
- The importance of moment generating function lies in the fact that, a probability distribution can be completely determined if its moments are known.
- The moment generating function of a random variable X is given by, , t > 0.
- A random experiment whose outcomes are classified into two categories, called 'success' and 'failure' is called a Bernoulli trials.
- The p. m. f. of binomial distribution with parameters and , = 0, 1, 2, . . ., n.
- The mean of binomial distribution is np and varience is npq.
- For a binomial distribution, mean is always greater than variance.
- Sum of independent binomial variates is also a binomial variate.
- Poisson distribution can be derived as a limiting case of binomial distribution when, , and
- The p. m. f. of poisson distribution with parameter , , = 0, 1, 2, ... ,
- Poisson distribution plays an important role in the discription of random phenomena in a continuous interval of any kind. For example, telephone calls at a telephone exchange within a specified period of time.
- Sum of independent poisson variates is also a poisson variate.

8.7 FURTHER READING

1) Das, K. K, Bhattacharjee, D (2008). A Treatise on statistical Inference and Distributions. Asian Books Pvt. Ltd.

2) Gupta, S. C. and Kapoor, V. K. (2002). Fundementals of Mathematical statistics; Sultan Chand & Sons.

3) Goon, A. M. Gupta. M. P and Dasgupta, B (2002). Fundementals of statistics (Vol I), Seventh Revised Edition; The World Press Pvt. Ltd. Kolkata.

4) Saxena, H. C. (1994). Probability and its application, sixth edition; S. Chand and Company, New Delhi.

5) Singh, R (2012). An Introduction to Probability and Probability Distributions, Books and Ellied (P) Ltd.

8.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True, (ii) False, (iii) False, (iv) True.

Ans to Q No 2: We know that if X ~ B (n, p) then E(x) = np.

Here, n = 12,

Ans to Q No 3: Here, given, mean = 5, s. d. = 3

variance = 9

But we know that for a binomial distribution, mean is always greater than variance. Hence, the above statement is false.

Ans to Q No 4: The parameters of a binomial distribution are n and p.

Ans to Q No 5: Given, , = 0, 1, 2, ... , 10.

Here, n = 10,

Hence, variance =

Ans to Q No 6: A particular trial having two outcomes only, viz success and failure occurring with probability p and (1–p) respectively is called a Bernoulli trial. e.g. tossing of a coin.

A random variable X is said to follow Beruoulli distribution if its p.m.f. is given by,

.

p is the only parameter of this distribution

Mean = E(x) = 0. (1–p) + 1 . p = p.

$V(x) = E(x^2) – \{E(x)^2\}$.

$E(x^2) = 0^2 . (x–p) + 1^2 . p = p$

$V(x) = p – p^2 = p (1–p)$.

Ans to Q No 7: Consider n independent Bernoulli trials each having two and two outcomes only called success and failure, Probability of success is p and the probability of failure is (1 – p) = q.

The probability of x successes and (n–x) failures in n independent trials in a specific order, say, s s s... f.f.f is given by,

P (s. s. . . . s. f . f. . . . f) = p(s) p(s) s . . . p(s) p(f) . . . p(f)

trials are independent

= p . p . . . p q q . . . q

x times (n-x) times

$= p^x q^{n-x}$.

But, x successes in n trails can occur in

mutually exclusive and the probability for each of these is $p^x q^{n-x}$. Hence by addition theorem of probability,

, = 0, 1, 2, . . ., n.

Ans to Q No 8: (i) equal, (ii) 2, (iii) 1.5, (iv)

(v) poisson.

Ans to Q No 9: Poisson distribution is also known as probability distribution of rare events, because poisson distribution is a positivaly skewed distribution, i.e probabilities tend to zero for large number of occurrences. For these poisson distribution is also known as probability distribution of rare events.

Ans to Q No 10: Some examples of poisson distribution are,

(i) number of deaths from suicide reported in a particular city.

(ii) number of defictive items in a lot of manufactures items.

(iii) number of printing mistakes in each page of a book.

8.9 MODEL QUESTIONS

Q 1: Define moment generating function and state its properties.

Q 2: Write a note on binomial distribution.

Q 3: Define a binomial distribution distribution and also obtain its mean and variance.

Q 4: Obtain poisson distribution as a limiting case of binomial distribution.

Q 5: If X follows poisson distribution with V(

Q 6: Define poisson distribution, state the conditions under which the binomial p.m.f. can be approximated by the poisson p.m.f.

Q 7: Obtain m.g.f. of poisson distribution and hence find its mean.

*** ***** ***

UNIT 9: THEORETICAL PROBABILITY DISTRIBUTIONS (CONTINUOUS VARIABLE)

UNIT STRUCTURE

9.1 Learning Objectives

9.2 Introduction

9.3 Normal Distribution

- 9.3.1 Definition
- 9.3.2 Standard Normal Variate
- 9.3.3 Properties of Normal Distribution
- 9.3.4 Area under Normal Probability Curve
- 9.3.5 Importance of Normal Distribution

9.4 Weak Law of Large Numbers

- 9.4.1 Statement
- 9.4.2 Application of Weak Law of Large Numbers

9.5 Central Limit Theorem

- 9.5.1 Statement
- 9.5.2 Application of Central Limit Theorem

9.6 Let Us Sum Up

9.7 Further Reading

9.8 Answers To Check Your Progress

9.9 Model Questions

9.1 LEARNING OBJECTIVES

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

- discuss about the shape and form of normal distribution, the most important theoretical distribution for continuous variable
- convert a normal variate to a standard normal variate and know its distribution
- understand how normal distribution can be obtained as a limiting case of binomial and poisson distribution
- find the area under the standard normal curve
- explain the important properties of normal distribution regarding its

moments, points of inflexion etc.

- have concept of weak law of large numbers and its application in practical field
- understand the importance of 'Central Limit Theorem' which gives normal distribution the central place in the theory of sampling.

9.2 INTRODUCTION

The binomial and poisson distribution studied in the previous unit are the most useful theoritical distributions for discrete variables, i.e. they relate to the occurrence of distinct events. In order to have a mathematical distribution suitable for dealing with quantities whose magnitudes vary continuously, a continuous distribution is needed. The normal distribution that we are going to discuss in the subsequent section of this unit is the most important theoretical distribution for continuous variable.

So normal distribution is the most widely known and used continuous distribution for data analysis. It is the most important distribution in applied statistics. Many statistical data concerning business and economic problems are displaced in the form of normal distribution. It has got this importance in statistics, because, most of the useful distributions in theoretical statistics, whether discrete or continuous tend to normal distribution when the number of observations are large. For example, normal distribution is an approximation to binomial distribution when the number of independent Bernoulli trials are very large and neither p nor q is very small. Again normal distribution can be obtained as a limiting case of poisson distribution when the mean of the distribution is very large.

Again in section 9.4 of this unit we shall learn about Weak Law of Large Numbers which tells us when a statistic which itself is a random variable will converge in probability to the parameter.

Again it is believed that the main reason for the widespread use of normal distribution in the area of statistical inference is the 'Central Limit Theorem' which we shall discuss in section 9.5 of this unit. According to this theorem the sample mean $\bar{x}$ from a population will approximate a normal distribution when the sample size n is quite large.

9.3 NORMAL DISTRIBUTION

The normal distribution was first discovered in 1733 by De Moivre, a French Mathematican , who obtained normal distribution as a limiting case of binomial distribution and applied it in the game of chance. But it was mainly known to Laplace and Gauss who made reference to it in the beginning of the 19th century.

If we consider a continuous random variable, say, heights of individuals or the yearly rainfall data over a long period of time and plot their probability density functions, we notice one interesting fact : most of those density curves resemble one another in shape, which is bell shapped symmetrical unimodal curve with most densites concentrated near the mode and very little density lying towards the two extremes. Thus most of the frequency, distribution, specially natural phenomena vary in accordance with this curve, called the normal curve, which is an indispensable tool for modern statistics. Let us now define a normal distribution as follows :

9.3.1 Definition

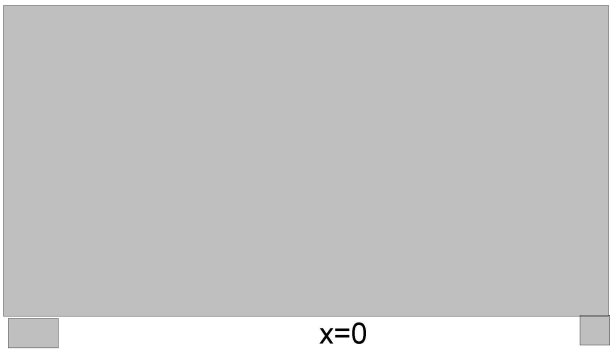
A continuous random variable X is said to have a normal distribution with parameter and

by.

; and

Note : A random variable X with mean and s.d. following normal probability law is expressed as

The probability curve of normal distribution is known as normal curve. This curve is bell shapped, symmetrical about mean two tails axtends to infinity on either side.



Normal curve

9.3.2 Standard Normal Variate

If X is normally distributed with mean and standard deviation , then the random variable

variate and the corresponding distribution is called standard normal distribution. If has the density function ,

also follows normal distribution with

E(z) = 0, v(z) = 1.

i.e. Z ~ N (0, 1).

The standard normal curve will look like the ordinary normal curve, only the diffirence is that is symmetrical about z = 0.



Standard normal curve.

9.3.3 Properties of Normal Distribution

A study of the normal distribution reveals the undermentioned significant propertics :

(i) The normal probability curve is bell shapped, symmetrical, unimodal curve. This means that the curve is single peaked and the distribution of frequencies on either side of the mean are same.

(ii) Mean, median and mode co-incide,

i.e.mean=median=mode=

(iii) For normal distribution, skewness = 0

kurtosis = 3

(iv) Sum of independent normal variates is also a normal variate.

(v) All moments of odd order are zero.

i.e.

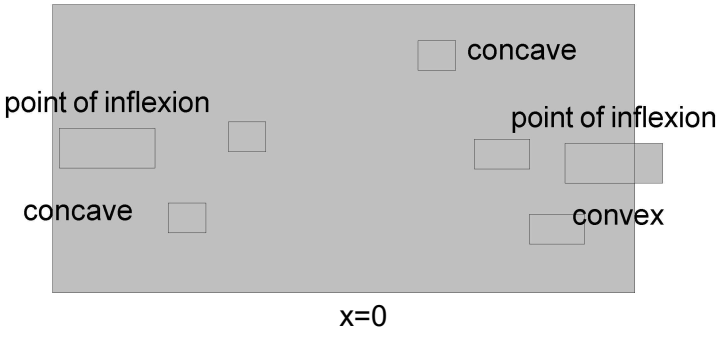
And moments of even order are given by the expression

; r = 1, 2, 3, ...

(i) Quartiles are equidistant from median.

(ii) The points and

inflexion. At these points the curve changes its curvature (i,e changes from convex to concave and concave to convex respectively) and are necessary for finding the shape of the curve.



(viii) The mean deviation is

standard diviation. i.e.

or, M. D = .7979

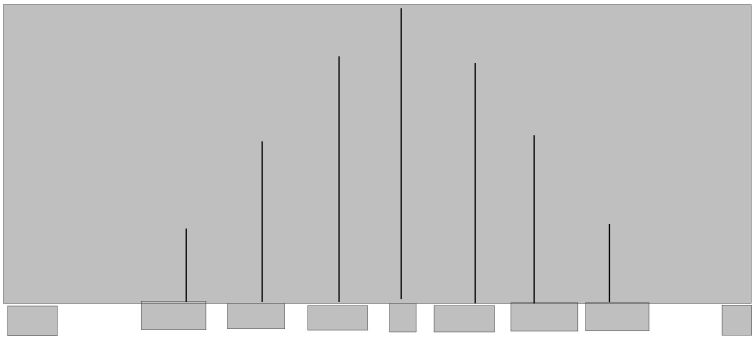
(ix) The area under normal curve is distributed as follows :

(a) covers 68.27% area, 34.135% area will be on either side of the mean.

(b)

(c)

This distribution of areas can be shown in the following diagram.



68.27%

98.45%

99.73%

Note : Let us now see how the discrete distributions discussed so far like binomial and poisson distributions tend to normal distribution.

(i) If X follows binomial distribution with parameters n and p, then it will tend to normal distribution if the following conditions are fulfilled.

(a) the number of independent Bernoulli trials is very large, i. e. .

(b) neither p nor q is very small.

(ii) If X follows poisson distribution with parameter tend to normal distribution if the mean of the distribution is very large, i.e.

9.3.4 Area under Normal Probability Curve

We know that p.d.f. of standard normal variate

,

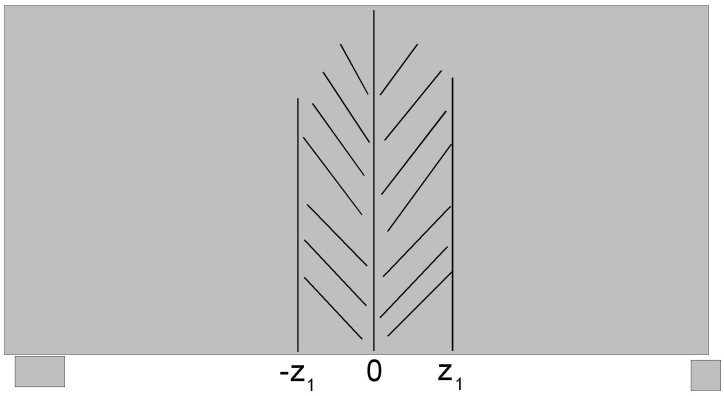
The area under this curve is equal to 1. In any problem, if we are interested in determining the area under normal curve we follow the following steps :

(i) At first we transform the given normal variate X to standard normal variate Z by the transformation

(ii) We use the table for 'Area under standard normal curve' given at the end of this book. Since in the table areas are given from Z = 0 to a particular positive quantity, so somehow or otherwise we transform the given area to that form.

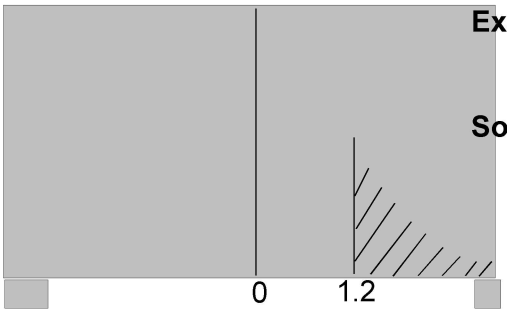
(iii) Since the normal curve is symmetrical, the probability that Z lies between, say, $-Z_1$ to 0 is equal to the probability that Z lies between 0 to Z_1.

i.e. $P(-z_1 < z < 0) = P(0 < z < Z_1)$



(iv)

Now we shall try to find the area under normal probability curve in the following examples.



Example 1: If X ~ N(66, 5), find

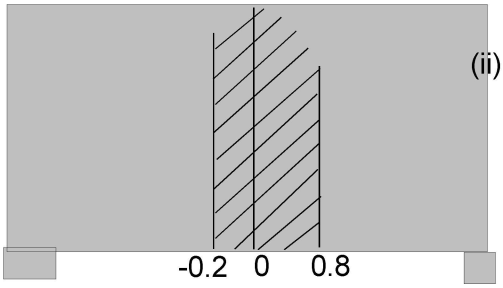
(i)

Sol^{n} : (i)

[shaded area is the requied area]

= 0.5 – 0.3849

= 0.115.



(ii)

= P(–0.2 < z < 0.8)

= P(–0.2 < z < 0) + P(0 < z < 0.8)

=P(0 < z < 0.2) + P(0 < z < 0.8)

= 0.0793 + 0.2881 [from table)

= 0.3674

Example 2: How many male workers in a factory have a salary between (i) Rs. 400 and Rs. 680, (ii) more then 720 if mean salary is Rs. 500 and s.d. is Rs. 100, and the salary of the workers assumed to be normally distributed. The number of workers in the family is 10,000.

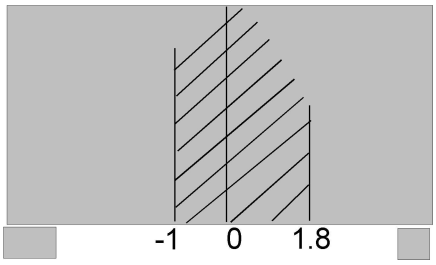
Solution: Let us assume that the random variable X denotes the salary of the workers.

We are given X ~ N (500, 100)

(i) Here we are to find the probability that the salary of the workers is between Rs. 400 and Rs. 680, i.e. we need,

P (400 < X < 680)

[transforming X to Z]



= P(–1 < z < 1.8)

= P (–1 < z < 0) + P (0 < z < 1.8)

= P (0 < z < 1) + P (0 < z < 1.8)

= 0.3413 + 0.4641

= 0.8054

So the number of male workers getting salary between Rs. 400 and Rs. 680 is,

0.8054 x 10,000

= 8054.

(ii) Here we want P(X > 720)






= 0.5–0.4861

= 0.0139.

Hence the number of male workers earning salary more than Rs. 720 is 0.0139 x 10,000

= 139.

Example 3: 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days.

(i) How many bulbs will expire in less than 100 days.

(ii) If it is decided to replace all the bulbs together, what interval should be allowed between replacements if not more than 10% should expire before replacement.

Solution: Let x devote the mean life of light bulbs, given that



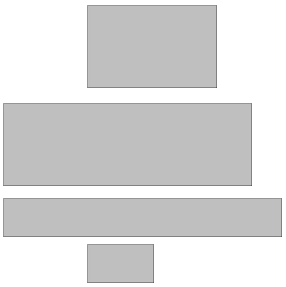
(i) To find the number of bulbs that expire in less than 100 days, we first find,







Hence the number of bulbs that should be replaced before 100 days is 1000 x 0.1587
= 159 (approx.)
(ii) The value of z corresponding to an area 0.4 (0.5-0.1) under the normal curve is 1.28.



Hence the bulbs will have to be replaced after 94 days.

9.3.5 Importance of Normal Distribution

The normal distribution has long occupied a central place in the theory of statistics. Its importance will be clear from the following points.

(i) As n increases, the normal distribution serves as a good approximation to many discrete distributions, such as binomial or poisson distribution.

(ii) One of the main reason behind the extreme use and application of Normal distribution is central Limit Theorem According to this theorem as sample size n increases, the distribution of mean

population approaches normal distribution.

(iii) The proof of test of significance in sampling theory on the fundemental assumption that the population from which sample has been drawn is normal.

(iv) The normal distribution is used extensively in statistical quality

control in setting up contral limits.

(v) The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate. The normal distribution is reasonably close to many distributions of the humped type. If, therefore, we are ignorant about the exact nature of the distribution, we may assume as a first approximation that the distribution is normal and see what this assumption leads us to.



CHECK YOUR PROGRESS

Q 1: State whether the following statements are true (T) or false (F).

(i) For a normal distribution

(ii) Normal probability curve is a asymmetrical curve. (T/F)

(iii) Normal distribution can be obtained as a limiting form of binomial distribution when

Q 2: Fill in the blanks.

(i) The probability distribution for which mean = median = mode is known as ..... distribution. (normal/binomial/ poisson).

(ii) If X ~ N , then (.9545/.9973)

(iii) The points of inflexion of normal distribution are ..... .

Q 3: Discuss briefly about the shape of the normal curve.

Q 4: Write four properties of normal distribution.

Q 5: There are six hundred students in the post graduate classes of a university and the probability for any student to need a copy of a particular book from the university library on any day is 0.05. How many copiess of that particular book should be kept in the university library so that the probability may be greater than 0.9 that none of the students needing a copy from the library has to come back disappionted. (use normal approximation to binomial distribution).

9.4 WEAK LAW OF LARGE NUMBERS

Suppose be the density function of a random variable X. If is known, then the density function will be completely specified and no interference from sample to the population will be necessary to consider a random sample which can tell us something about the unknown parameter.

Weak law of large numbers tells us that a positive integer n can be determined such that if a random sample of size n or large is taken from a population with density , with , the probability that the sample mean will deviate from

to 1. We now state the Weak Law of Large Numbers as follows.

9.4.1 Statement

Let $X_1, X_2, ..., X_n$ be a sequence of random variables having expectations

$B_n = var\ (x_1 + x_2 + ... + x_n) <$

If as , then, given any two positive quantitices , however small may be, we can final an depending on such that,

for all

Note : For existence of Weak Law of Large Numbers the following three conditions should be satisfied by the random variable X_i, i = 1, 2, ..., n. They are,

(i) $E(X_i)$ exists and is finite.

(ii) $B_n = V(x_1 + x_2 + ... + x_n)$ exists.

(iii) as

The condition (i) is necessary, without which the law cannot be stated. Condition (ii) and (iii) are not necessary. However condition (iii) is a sufficient condition.

9.4.2 Application of Weak Law of Large Numbers

Example 4: Let $\{X_n\}$ is a sequence of mutually independent and identically distributed random variables with mean [redacted] variance. If $S_n = X_1 + X_2 + ... + X_n$, prove that the law of large numbers doesnot hold for the sequence $\{S_n\}$.

Solution: Consider the variables $S_1, S_2, ... S_n$.

Now, [redacted] $_1, S_2, ... S_n)$

$= V\{x_1+(x_1+x_2)+(x_1+x_2+ x_3) +... (x_1 + x_2 + ... + x_n)\}$

$= V\{nx_1+(n–1)x_2+(n–2)x_3 +...+ 2x_{n-1} + x_n\}$

$= n^2v(x_1)+(n–1)^2 v(x_2)+(n–2)^2 v(x_2) +...+ 2^2 v(x_{n-1})+ 1^2 v(x_n)$

[Covariance terms vanish as $x_1, x_2, ..., x_n$ are independent r. v's and $v(ax) = a^2 v(x)$]

Let, $v(x_i)$ = [redacted]

[redacted]

Hence Law of Large Numbers doesnot hold for the sequence $\{S_n\}$.

9.5 CENTRAL LIMIT THEOREM

A central limit theorem asserts that under quite general conditions, the standardized sum of independent normal variates has an approximately normal distribution. It occupies a place of honour in probability theory acquired by the fruitfull role which it played and still plays in the development of theory applications. It is the central limit theorem which gives the normal distribution it's central place in the theory of sampling, since many important problems can be solved by this single pattern of sampling variability. As a result the work on statistical inference is made simpler. Now we want to state central limit theorem as a generalization of laws of large numbers as follows :

9.5.1 Statement

If $x_1, x_2, ..., x_n$ are n independent random variables with $E(x_i)$ = $_i$) =

certain general conditions the random variable $S_n = x_1 + x_2 + ... + x_n$ is asymptotically normal with mean and variance

and

Note : This theorem was first stated by Laplace in 1812 and a rigorous proof under certain general conditions was given by Leapounoff in 1901. Below we shall consider some particular cases of this general central limit theorem.

De- Moivre Laplace Central Limit Theorem :

Statement : If

then the distribution of $S_n = x_1 + x_2 + ... + x_n$, where x_i's (i = 1, 2, ...n) are independent, is asymptotically normal as

Lindeberg - Levy Central Limit Theorem :

Statement : If $x_1, x_2, ..., x_n$ are independently and identically distributed random variables with

i = 1, 2, ..., n.

then the sum $S_n = x_1 + x_2 + ... + x_n$ is asymptotically normal with mean and variance

Note : CLT can be stated in another form as follows :

If $x_1, x_2, ..., x_n$ are i. i .d with mean and variance

and $S_n = x_1 + x_2 + ... + x_n$, then

9.5.2 Application of Central Limit Theorem

Below are given some applications of CLT in commonly used distribution of random variables.

1. If $x_1, x_2, ..., x_n$ are i i d binomial variates with parameters (k, p), $0< p < 1$, then,

$S_n = x_1 + x_2 + ... + x_n$. as

Proof : Now,

Hence by CLT we have,



CHECK YOUR PROGRESS

Q 6: State true or false.

(i) The law of large number tells us that a statistic will converse to a parameter in some sense. (T/F).

(ii) The central limit theorem are named so because they are central to the distribution theory necessary for statistical inference. (T/F)

Q 7: Let x_i assume values i and –i with equal probability show that WLLN cannot be applied to independent variables $x_1, x_2, ..., x_n$.



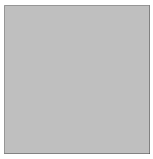
9.6 LET US SUM UP

- Normal distribution is the most widely known and used continuous distribution for data analysis.

- p.d.f. of normal distribution is,

, 

-  and 
  mean and standard deviation of this distribution.
- Mean, median and mode occur at the same point, i.e mean = median = mode.
- Quartils are equidistance from median.
- It is a symmetrical distribution.
- skewness = 0, kurtosis = 3.
- Binomial distribution tends to Normal distribution if,
  (i) n, the number of independent Bernoulli trials is very large.
  (ii) neither p nor q is very small.
- Poisson distribution with parameter 
  .
- To find the area under normal curve we have to transform the random variable x to standard normal variate z.
- Central limit theorem establishes conditions under which sum of independent random variables are asymptolically normally distributed.



9.7 FURTHER READING

1) Gupta, S. C. Kapoor, V.K., Fundementals of Mathematical statistics (2002). Sultan Chand and Sons.

2) Goon, A. M, Gupta, M. P. and Dasgupta, B, Fundementals of statistics, Vol I, Seventh Revised Edition (2002). The World Press Pvt. Ltd, Kolkata.

3) Das, K.K and Bhattacharjee, D; A Treatise on statistical Inference and Distribution (2008). Asian Books Private Limited.

4) Kakoty, S. C.; Mathematical Statistics : Theory and applications (2003); Kasturba Prakashan, Dibrugarh, Assam.

5) Saxena, H.C.; Probability and its application (1994); S. Chand and Company, New Delhi.

9.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True, (ii) False, (iii) False.

Ans to Q No 2: (i) Normal, (ii) .9545, (iii)

Ans to Q No 3: The probability curve of Normal distribution is known as normal curve. The curve is symmetrical about its mean shapped and extends to infinity on either side. The height of the normal curve is at its maximum at the mean. That is there is only one maximum point of the curve and the height declines as we go on either direction from mean. The curve is asymptotic to the base.

Ans to Q No 4: The four important properties of normal distribution are,

(i) the normal curve is bell shapped and symmetrical about mean.

(ii) for normal distribution, mean = median = mode.

(iii) skewness = 0, kurtosis = 3

(iv) the linear combination of independent normal variates is also a normal variate.

Ans to Q No 5: Let x devote the number of copies of the particular book.

p(x) = ^{x} q^{n-x}, x = 0, 1, 2, ... n.

We are given, n = 600, p = 0.05.

[^{n} is np]

= 600 x 0.05 = 3

varianes =

We need x = x, such that

P(X < x) > 0.90

where

$z > 1.28$ from normal probability table.

Hence the university library should keep at least 37 copies of that book.

Ans to Q No 6: (i) True, (ii) True.

Ans to Q No 7: Given that, $P(x_i = i) = p$, $P(x_i = -i) = p$

$E(x_i) = i.p + (-i).P = 0$

$E(x_i^2) = i^2.p + (-i)^2.p = 2\,i^2p.$

$V(x_i) = E(x_i^2) - \{E(x_i)\}^2 = 2\,i^2p.$

$B_n = v(x_1 + x_2 + ... + x_n)$

$= V(x_1)+v(x_2) + ... + v(x_n)$

$= 2.1^2\,p + 2.2^2p + ... + 2.n^2\,p$

$= 2p(1^2 + 2^2 + ... + n^2)$

=

=

Now,

Weak Law of Large Numbers can not be applied in this case.

9.9 MODEL QUESTIONS

Q 1: Under what conditions binomial distribution can be approximated to normal distribution.

Q 2: Define normal probability distribution. If the mean of a normal distribution is and variance is , what are its median, mode. and

Q 3: Write a note on the salient features of a normal distribution.

Q 4: Define a standard normal variate and obtain its m.g.f.

Q 5: If x follows normal distribution with mean and s.d. distribution of Y = a + bX.

Q 6: The monthly income of a group of 10,000 persons found to be normally distributed with mean Rs. 750 and s.d. Rs. 50. Show that of this group, about 95% had income exceeding Rs. 668 and only 5% had income exceeding Rs. 832.

Q 7: State Weak Law of Large Numbers.

Q 8: Let $\{X_n\}$ be a sequence of mutually independent random variables such that

*** ***** ***

UNIT 10: SAMPLING AND POPULATION

UNIT STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.3 Concept of Population and Sample

- 10.3.1 Basic Principles of Sample Survey
- 10.3.2 Advantages of Sampling over Complete Enumeration
- 10.3.3 Different Steps in a Sample Survey
- 10.3.4 Errors in Surveys (Non-Sampling and Sampling Errors)
- 10.3.5 Types of Sampling
- 10.3.6 Selection of Random Samples

10.4 Simple Random Sampling

- 10.4.1 Estimate of Population Mean and Population Total
- 10.4.2 Illustrated Examples

10.5 Stratified Random Sampling

- 10.5.1 Advantage and Disadvantages of Stratified Random Sampling
- 10.5.2 Allocation in Stratified Random Sampling

10.6 Systematic Random Sampling

- 10.6.1 Linear Systematic Sampling
- 10.6.2 Circular Systematic Sampling
- 10.6.3 Advantages and Disadvantages of Systematic Sampling
- 10.6.4 Uses of Systematic Sampling

10.7 Cluster Sampling

- 10.7.1 Advantages and Disadvanges of Cluster Sampling
- 10.7.2 Optimum Cluster size

10.8 Let Us Sum Up

10.9 Further Reading

10.10 Answers To Check Your Progress

10.11 Model Questions

10.1 LEARNING OBJECTIVES

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

- distinguish between a sample and population
- state the difference between parameter and a statistic
- understand why sample surveys are adopted instead of complete enumeration in most cases
- acquire knowledge about the types of errors that one encounters while adopting sample surveys
- distinguish between random sampling and purposive sampling
- draw samples using simple random sampling technique
- discuss the advantages and disadvantages of using stratified random sampling from a heteregeneous population
- explain the process of linear systematic and circular systematic sampling and their relative merits and demerits
- evaluate the advantage of drawing cluster sampling and how to decide optimum cluster size.

10.2 INTRODUCTION

The use of sampling in making inference about a population is a practice that we are doing as a part of our everyday's life from times unknown. When we have to make an inference about a large lot and we are not in a position to examine each and every individual of the lot, we examine only a few members of the lot and make decission about the whole lot. For example a person wanting to purchase some rice will examine a handful of rice from a bagful of rice and on that basis make decisions about the whole bag of rice. Similarly in industry quality of a product is assured through sampling, the public opinion is ascertanied through opinion polls based on sampling. In all these cases inference is drawn about the whole from the knowledge of a part.

So what we see is that most of the enquiries that we make in practice are sampling enquires, sampling is inevitable because our resources are always limited, be it money or man hours. It is also preferred because it has

practical conveniencs and there are many situation where complete count is impossible.

In this unit we shall discuss how to use different sampling techniques to draw our samples according to the type of population and purpose of study. In section 10.3 we shall learn about basic concepts of sampling, its benifits, errors etc. In the subsequent, chapters different sampling techniquies like simple random sampling, stratified random sampling, systematic sampling and cluster samplings are discussed along with their relative advantages and disadvantages.

10.3 CONCEPT OF POPULATION AND SAMPLE

Population : The dictionary meaning of the word population is the number of people residing in a geographical area. But in statistics the word population is used in much wider sense. Thus population may be defined as a group of individuals with respect to certain characteristic under study. For example if we want to measure the average height of students of a college, then the students of that college will constitute a population. Population may be finite or infinite.

Sample : The part of the population which is selected from it in order to study its characteristic is called a sample. For example if we want to measure the standard of living of the people in a town of population 10,000, there if we estimate this characterestic from a properly selected 500 members of the population, then these 500 members of the town will constitute the sample.

Complete enumeration or census : It refers to the enumeration of each and every unit of the population. We all know about the census the government of India conducts in every ten years.

Sample survey : When we enumerate units belonging to the sample only, it is sample survey.

Parameter : The basic objective of any sample survey is to know the characteristic of the sample by examining only a part of it, by studying the characteristic of the population. So the term parameter is used to denote the characteristic of the population. For example population mean, population

variances are parameters. Parametrs are constant quantities.

Statistic : Values obtained from the study of the sample are known as statistic, sample mean, sample variance are the examples of statistic. statistics are variable quantities.

Sampling unit : The individuals that are to be sampled from the population to form the sample is called the sampling units. For example if we want to measure the average income of the families in a city, then each household may be regarded as sampling unit, but if we want to know the average income of individuals in the same city, then the individuals will be the sampling units.

Sampling frame : Sampling frame is the complete list of sampling units. Such sampling frames are difficult to construct and even difficult to keep it update. Simplest example of sampling frame is the voters' list.

Sampling error : It is the difference between the parameter and its estimate from the random sample, It is ineviable because only a part of the population is selected.

Unbiased estimate : Let  sample values . This statistic is called an unbiased estimate of the parameter , if E (T) =

Sampling variances : The variance of a statistic is called the sampling variance.

Standard error : The positive square root of sampling variance is the standard error.

10.3.1 Basic Principles of Sample Survey

There are two important laws on which the theory of sampling is based. They are principles of validity and optimization.

Principle of validity means that the sampling design should be such that valid estimates and tests about the parameter can be obtained.

Again optimization means either maximization or minimization. According to this principle such sampling design should

be preferred which gives maximum efficiency for a given cost of survey or the minimum cost for a specified level of efficiency.

10.3.2 Advantages of Sampling over Complete Enumeration

In most cases sample survey is preferred over complete enumeration because of the following conditions :

- In sample survey since only a part of the population is selected, there is reduction of cost either in terms of money or in terms of man hours. Although the cost per individual may be larger in sample survey, the total cost is yet expected to be smaller. In most of our cases resources are limited and the result of the survey is required within a specifeed times, it is always imperative to adopt sample survey.
- In certain types of enquirey highly trained personnel or specialised equipments must be used for collection of data. This is not possible in census method because a large group of investigators are to be employed and it will be highly expensive to train such a large group. So for more accurate results sampling is the only way out.
- In sample survey sampling errors cannot be avoided. On the otherhand though sampling error is absent in complete enumeration, large magnitude of non sampling errors crop in, whose magnitude can not be precisely determined and they greatly influence the accuracy of results. However sampling errors can be theoritically calculated and hence its effect on the accuracy of results can be minimized.
- If the enumeration is by its nature destructive, then complete enumeration is not feasible. For example if we want to examine the quality of sweets in a sweet shop, we have to recourse to a sample. Also in case of infinite or hypothetical population sampling is the only course available.

10.3.3 Different Steps in a Sample Survey

Conducting a large scale sample survey involves three main stages : (a) Planning stage, (b) Execution stage, (c) Analysis and reporting stage.

The planning stage consists of the following steps :

- **Defining the objectives :** The objectives of the survey should be clearly stated. The planner should also take into account the available resources, the time limit within which the study should be finished, the degree of accuracy required etc.
- **Defining the population :** The population along with geographical and demographical boundaries must be clearly defined so that no ambiguity arise later on regarding the coverage of the study.
- **Determination of the data to be collected :** This should be done in accordance with the objective of the survey. Once the decision is made about the nature of the data to be collected, one must prepare the questionnaire or a schedue for collection of data.
- **Method of collection of data :** Decision is to be made whether the interview method or the mailed questionnaire method is to be adopted. Though the second method is less costly there is large scope for non-response and is praticable only among the educated masses.
- **Choices of sampling unit :** The sampling unit is the ultimate unit to be sampled for the purpose of the survey. For example depending on the requirement of the survey in a socio-economic survey the ultimate sampling unit may be a family or a member of a family.
- **Getting a sampling frame :** Once the sampling unit is decided, the next step is to see whether the sampling frames of all the units of the population is available. An example of sampling frame can be the list of households from population census of a region.

- Designing the survey – This is the most important part of planning stage. At this stage decision is made about the type of sampling is to be used and about the need of pilot survey.
- Training of personnel – The investigators and the supervisor should be properly trained before the survey is undertaken.

The **Execution Stage** involves identification of the sampled individual in the field and filling up questionnaire.

The first step in **Analysis and Reporting stage** is to edit completed questionnaires deleting data that are obviousely erroneous. The next step is to tabulate the data and then these tables are utilised for deriving necessary estimatis of population characteristics.

The next step is report writing. The report should incorporate a detailed statement regarding all the stages of the survey and all the statistical informations collected in a neat tabular form. It is necessary to report the amount of error to be expected in the estimates.

At the completion of the survey, the informations should be properly stored for possible use in designing future surveys.

10.3.4 Errors in Surveys (Non-Sampling and Sampling Errors)

To appreciate the need for sample surveys, it is necessary to understand the role of sampling and non-sampling errors in sample survey and complete enumeration. The errors arising due to drawing inference about the population on the basis of sample is known as sampling errors, clearly sampling error is non-existent in complete enumeration, since the whole population is enumerated. Howerver some other kind of error crop-in, at the stage of ascertainment and, precesing of data, which are termed as non-sampling errors and are common in both complete enumeration and sample surveys. We shall discuss about these two types of errors one by one.

Non sampling errors : As discussed above non-sampling errors are present in both complete enumeration as well as sample surveys. Below we shall study different types of non-sampling errors.

- **Respons errors:** These errors have their origin in the responses or informations provided by the respondents. For example wrong answers arising from pride, called error due to prestige, by virtue of which one may overstate ones educational qualification, or occupation or understate one's age. Sometimes one may provide wrong answers for the protection of self interest. For example people generally tend to make an understatement of their incomes and overstatement of their expenses. Sometimes respondent may misunderstands a particular question and hence may provide wrong answers un willingly.
- **Errors arising from non-response:** Often at the times of collection of data some respondents are not found at home even after repeated calls, or he or she simply refuses to furnish information. As a result of these non-responses the effective simple size decreases and certain characteristic of the population gets excluded, which gives rise to error.
- **Observational error :** This type of error occurs when the variate value is obtained by observation, because of the influence of somes psychological factors. For example in case of eye condition of crop condition, the crop reporters invariably understimate the result.
- **Interviewer error :**Sometimes the suggestions from the interviewer may influence the beliefs and prejudices of the respondents in interpreting some questions and thereby leading to some error.

Sampling Errors :

These errors have their origin in sampling and are absent in complete enumeration. Some errors of this type are mentioned below.

- **Error due to defective sampling technique :** Samples can generally be drawn in two ways– (a) Random sampling and (b) Purposive sampling.

  Random sampling is free from any type of errors because of the random nature of sampling. But in case of purposive sampling the sampler deliberately tries to select a representative sample to establish his own belief, which obviousey gives rise to errors. For example in a socio-economic survey of the city of Guwahati, the sampler may select only the people belonging to upper class if the investigator wants to establish that the people of the city of Guwahati are very rich.

- **Error due to substitution :** Investigators often substitute one convenient member of the population when difficulties arise in enumerating another, which may lead to errors as the characteristics of the two persons may be quite different.

- **Error due to wrong choice of statistic :** Such errors may arise when the investigator doesnot use appropriate statistic for the estimation of a particular parameter of the population. For example, we know that sample mean square is an unbiased estimate of the population variance, but one may use sample variance in place of sample mean square, leading to error.

10.3.5 Types of Sampling

Whenever we decide to take sample from a population it is necessary to choose some method in which sampling is to be done. There are various methods of choosing a sample from a population, depending on the nature of the data and the purpose of enquiry. However these various methods of sampling can be grouped under two broad heads : Probability sampling (also known as random sampling) and non-probability (non random) sampling.

Probability sampling : Probability sampling methods are these in which every item in the universe has a known chance or probability of being chosen for the sample. This implies that the selection of the sample item is independent of the person making

the study, i.e the items are chosen strictly at random.

Non Probability sampling : Non-probability sampling, also called 'Purposive' or 'judgement' sampling is the process where the sample is selected on the basis of experenice or according to the conveneince of the user. Hence unlike random sampling, here, the probability of selection of some of the units of the population are very high, and some are less. This method is not recommended for general use. However in case of experined investigator judgement sampling may give valuable results.



CHECK YOUR PROGRESS

Q 1: Fill in the blanks.

(i) If all the units of the population are enumerated then it is called ..... (sample/population)

(ii) A population consisting of unlimited number of units is called a ..... population. (finite/infinits)

Q 2: State whether the following statements are true or false.

(i) Sampling errors are present both in complete enumeration and sample survey. (T/F).

(ii) Judgement sampling is not recommended for general use. (T/F)

Q 3: In what situation sampling is inevitable.

Q 4: Mention in brief the objective of sampling.

Q 5: Distinguish between parameter and statistic.

10.3.6 Selection of Random Samples

Random samples are mainly selected either by (i) Lottery Method or (ii) Random number table method.

Lottery Method : In this method all the units of the population are numbered and the numbers are written on some cards which are identical in shape, size and colour. These cards are then placed in a jar and shuffled properly. A blind-fold selection is made from these cards untill the desired sample size is acheived. However shuffling

is necessary after each draw. This is a very reliable method of selection of random sample. But when the population is large this method is not practical.

Random number method : This is the most widely used method of selection of random samples. Here at first we number the population. Then we select at random any page of a random number table and pick up numbers in any row or column at random. Items corresponding to numbers thus selected constitute a random sample.

Note : Random number tables are so constructed that each of the digits 0, 1, 2, ..., 9 occurs independently and approximately with same frequency. Some commonly used such tables are 'Teppets Random Number Table'. 'Fisher and Yates Table' etc.

10.4 SIMPLE RANDOM SAMPLING

The simplest and the most commonly used type of probability sampling is simple random sampling. In this sampling each unit of the population has the same probability of being included in the sample. Simple random sampling may be of two types – simple random sampling with replacement and simple random sampling without replacement, as discussed below :

Simple random sampling with replacement (SRSWR) :

In this type of sampling the members of the sample are drawn one by one from the population and after each draw and the selected number is returned back to the population. As a result any member of the population may appear more than once in the sample. So if we want to draw a sample of size n from a population of size N, then each member of the population has the same probability of being selected and the total number of possible samples is N^n.

Simple random sampling without replacement (SRSWOR) :

In SRSWOR the sample units are drawn one by one and the members selected at any draw are not returned back to the population. Obviously in this case no member of the population can appear more than

once in the sample, i.e. the sample units are distinct. If we want to draw a sample of size n from a population of size N, the chance of selecting any unit at the first draw is , at the 2nd draw it is , at the third draw and so on. In SRSWOR the total number of possible samples are N_{c_n}.

10.4.1 Estimate of Population Mean and Population Total

Before going to find the estimate of population mean and population total we shall have to know some important notations as given below.

The N units of the population are denoted by $X_1, X_2, ..., X_N$

The n units of the sample are denoted by

is the population mean.

is the population variance.

is the population mean square.

is the sample mean.

is the sample variance.

is the sample mean square.

Using the above notations we can state a very important theorem as stated below.

Theorem. 10.4.1

Statement : In simple random sampling sample mean is an unbiased estimate of the population mean.

In symbol, we can write,

Again we know that population total is

Estimate of population total is

i.e.

Some useful Results :

(i) is the sampling fraction and is considered negligible if its valus is less than 0.05.

(ii)

ignored if N is vary large compared to n.

(iii) In SRSWR, the sampling variance of estimate of population mean is given by,

(iv) In SRSWOR, the sampling variance of estimate of population mean is given by,

(v)

Note : The sampling design having less variance is always more efficient. Hence SRSWOR is more efficient than SRSWR sampling design.

10.4.2 Illustrated Examples

Example 1: Suppose that 4 units of a population are $X_1 = 2$, $X_2 = 4$, $X_3 = 6$, and $X_4 = 8$. Draw all possible random samples of size 2 without replacement and calculate their means, show that .

Sol^n : Here we have N = population size = 4

n = sample size = 2

Total no. of samples drawn by SRSWOR is $4_{c_2} = 6$, as given below:

Serial No.	Sample	Sample mean
1	(2. 4)	=3
2	(2, 6)	4
3	(2, 8)	5
4	(4, 6)	5
5	(4, 8)	6
6	(6, 8)	7
Total		

.



CHECK YOUR PROGRESS

Q 6: State true or false.

(i) When every unit of the population has the same probability of coming to the sample it is called simple random sampling. (T/F)

(ii) In SRSWOR the same item may appear more than once in the sample.

Q 7: Fill in the blanks.

(i) = ...

(ii) In simple random sampling estimate of population total is .....

Q 8: Distinguish between SRSWR and SRSWOR.

10.5 STRATIFIED RANDOM SAMPLING

Till now we have discussed the technique of drawing simple random samples and also obtained estimates of the population parameters. But simple random sampling can give a representative sample only when the

population under consideration is homogeneous with respect to the characteristic under study. But in practice we come across populations which are not homogeneous in nature. The sample technique used for such a population is stratified random sampling which by using available information concerning the population, attempts to design a more efficient sample than obtained by simple random sampling technique. While applying stratified random sampling technique, we follow the steps given below:

(i) The population to be sampled is subdivided into groups called strata which are non-overlapping and collectively exhaustive.

(ii) A simple random sample is then chosen independently from each group, untill the required sample size is achieved.

However, for maximum efficiency the strata should be so formed that the units within a strata are as homogeneous as possible and a marked difference is possible between the strata.

For illustration suppose, Let our heterogeneous population of size N is divided into k strata, so that, there are

N_1 units in the 1^{st} stratum

N_2 units in the 2^{nd} stratum

. . . . . .

N_k units in the k^{th} stratum.

So that $N_1+N_2+.....+N_k = N$.

We now want to draw a sample of size n from this population so that we draw.

n_1 units from the 1^{st} stratum

n_2 units from the 2^{nd} stratum

. . . . . .

n_k units from the k^{th} stratum

such that $n_1+n_2+.....+n_k = n$. Only if the samples from each stratum are drawn using Simple Random Sampling then the method is called stratified random sampling.

10.5.1 Advantages and Disadvantages of Stratified Random Sampling

Advantages :

- This method ensures better representation of the characteristic of the population.
- Division of a big population into small sub-populations can bring administrative convenience. For example the agency conducting the survey may employ different field officers for different parts of the population.
- Sometimes different degree of accuracy is needed for different sections of the population, which may be possible only in stratified random sampling.
- In case of skewed data like income data this method is very effective.

Disadvantages :

- It is important to decide the basis of stratification well in advance, otherwise the results may be misleading.
- Each stratum must have, as far as possible homegeneous items, otherwise the results may not be reliable. If proper stratifiction is not done, the sample may have the effect of bias.
- Because of the likelihood that the stratified sample will be more widely distributed geographically than a random sample, cost per observation may be quite high.

10.5.2 Allocation in Stratified Random Sampling

In stratified random sampling, the allocation of sample to different strata is done by considering three factors, viz.

- the total number of units in the stratum,
- the variability within the stratum,
- the cost in taking observations per sampling unit in the stratum.

A good allocation is one where maximum precision is obtained with minimum resources. In otherwords, allocation of sample size

should be such as to minimize the budget for a given variance or minimize the variance for a fixed budget, so that there is most effective use of available resources.

There are three methods of allocation of sample size to different strata in stratified sampling technique. They are: Proportional allocation, Neyman allocation and Optimum allocation.

- **Proportional allocation :** This allocation was originally proposed by Bowley (1926). This procedure of allocation is very common in practice because of its simiplicity, where the size of the sample from each stratum varies with the size of the stratum. Larger the size of the stratum. Larger is the size of the sample selected from that stratum. In proportionall allocation, one samples each stratum in proportion to its relative weight.
- **Neyman allocation :** This allocation of total sample size to strata is called meximum variance allocation and is due to Neyman (1934). Here the allocation of samples among different strata is proportional to the product of the size and variability of the stratum.
- **Optimum allocation :** In this method of allocation of the sample size in the respective strata with a view to minimize variance for a specified cost of the survey or to minimize the cost for a specified value of the variance.

Illustrated example :

Example 2: The population of three towns are 40,000, 30,000 and 20,000 respectively. A stratified random sample of size 400 is to be drawn from this population using proportional allocation. Determine the sample size salected from each stratum.

Solution: Here, N_1= 40,000, N_2= 30,000, N_3= 20,000

$N = N_1+N_2+N_3 = 90,000$

sample size = n = 400

Now, n_1 = sample size from 1st stratum



Obveovity $n_1 + n_2 + n_3 = 400 = n$.



CHECK YOUR PROGRESS

Q 9: State wheter the following statements are true (T) or false (F).

(i) stratified sampling is preferred when the population is homogeneous. (T/F)

(ii) Simple Random sampling method is more efficient than stratified random sampling method (T/F)

10.6 SYSTEMATIC RANDOM SAMPLING

Till now we have discussed sampling techniques where sampling units were selected randomly. The sampling technique we are going to discuss now is also known as quasi-random sampling technique, because in this sampling technique only the first unit is selected at random. Once the interval starting point is determined the remainder of the items selected for the sample are predetermined by the sampling interval.

So a systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals untill the desired sample size is achived. This method is popularly used in those cases where a complete list of the population from which sample is to be drawn is available. The list may be prepared in alphabitical, geogrephical, numerical or some other order. The items are then serially numbered. The first item is selected at random, generally by using 'Lottery Method'. Subsequent items are selected by taking every k^{th} unit of the list, where k stands for sampling interval, i.e. 

where, n = sample size,

N = population size

Systematic sampling is simple and fool proof. We shall first discuss the sample selection procedures before going to discuss its advantages and desadvantages.

10.6.1 Linear Systematic Sampling

As discussed above, a common procedure is the linear systematic sampling scheme. We suppose that the N units of the population are linearly ordered from which a sample of size n is to be drawn. Here two cases may arise :

Case 1 : Let N be expressible in the form N = nk, where k is an integer, called the sampling interval.

In this procedure, a random number

the sample will consists of the unit and every k^{th} unit thereafter. Hence, under this procedure the sample comprises of the units .

The technique will generate k systematic samples with equal probability which may be shown in the following table.

Sample Number			
1	2		k
y_1	y_2		y_k
.	.		.
.	.		.
.	.		.

Case 2: N is not expressible in the form N = nk

Here the present sampling scheme will give rise to samples of unequal sizes. Here also we follow the same procedure with the

random selection of the 1st unit and selection of every k^{th} unit thereafter. Since in this case , k is taken as an integer nearest to

in some cases it may be less than n.

To make this concept more clear let us consider a population of 12 units from which we are to draw a systamatic sample of size 4.

i.e N = 12, n = 4

In this case N = nk where k = 3. Hence the possible samples will be,

	Selected sample
1	(1, 4, 7, 10)
2	(2, 5, 8, 11)
3	(3, 6, 9, 12)

Now, let N = 11, n = 4.

So

So the samples will be.

	Selected sample
1	(1, 4, 7, 10)
2	(2, 5, 8, 11)
3	(3, 6, 9)

i.e, when N is not a proper multiple of n all the samples are not of equal size. To overcome this difficulty of varying sample size we may use circular systematic sampling.

10.6.2 Circular Systematic Sampling

So, as discussed above the disadvantage of sample size being different from actual sample size required, when procedure is slightly modified by which a sample of constant size is always obtained. The procedure consists in selecting a unit by a random start, from 1 to N and thereafter selecting every k^{th} unit, k

being an integer nearest to 

of n units is obtained.

For illustration, let N = 11, n = 4, so that k = 3 (apprx.) So the possible samples are,

(1, 4, 7, 10), (2, 5, 8, 11), (3, 6, 9, 1), (4, 7, 10, 2), (5, 8, 11, 3), (6, 9, 1, 4), (7, 10, 2, 5), (8, 11, 3, 6), (9, 1, 4, 7), (10, 2, 5, 8), (11, 3, 6, 9).

10.6.3 Advantages and Disadvantages of Systematic Sampling

Advantages :

1) Systematic sampling is operationally more convenient than simple random sampling or stratified random sampling, because it requires less time and less work involved.
2) Systematic sampling may be more efficient than simple random sampling when the list from which the sampling units are drawn is arranged at random.

Disadvantages :

1) Systematic sampling becomes less representative if we are dealing with populations having 'hidden periodicities'.
2) If the population is ordered in a systematic way with respect to the characteristic under study, then it is possible that only certain types of items will be included in the sample, or at least more of certain types of items than others. For instance, in a study of workers wages the list may be such that every tenth worker on the list gets above Rs. 2,000 per month.

10.6.4 Uses of Systematic Sampling

Systematic sampling is operationally convenient and efficient in sampling some natural population like forest areas for estmating volume of timber, areas under different types of covers etc.



CHECK YOUR PROGRESS

Q 10: State whether the following statements area true (T) or false (F).

(i) When the population consists of units arranged at random one would prefer systematic sampling. (T/F)

(ii) When the population size is not a proper multiple of sample size, one should prefer circular systematic sampling rather than ordinary systematic sampling. (T/F)

Q 11: In what situation systematic sampling is preferred over other sampling techniques.

Q 12: Compare systematic sampling with stratified sampling technique.

10.7 CLUSTER SAMPLING

In random sampling we have seen that the population is divided into a finite number of distinct and identifiable elements known as sampling units. The smallest units into which the population is divided is called an element of the population. A group of such elements is called a cluster. So when the sampling unit is a cluster, rather than a single element of the population, the sampling procedure is known as cluster sampling.

Therefore, in cluster sampling the entire area containing the population is divided into some smaller area segments and each element of the population belongs to one and only one segment of the population. This procedure is also called ‘area sampling’.

Generally, in any sampling procedure the main problem lies in identification and location of an element. However once an element has been located, the time taken for surveying a few neighbouring elements is less. That is where the justification for formation of cluster lie. But as clusters are made-up of neigbouring elements, they may have similar characteristies. Hence it is always advisable to form small clusters and the number of clusters should be more.

After dividing the population into clusters, the required number of clusters are selected by equal or unequal probabilities of selection. Again all elements in selected clusters are enumerated.

10.7.1 Advantages and Disadvantages of Cluster Sampling

Advantages :

- Cluster sampling is less expensive and requires less times compared to other sampling techniques discussed so far. It is so because selection of neighbouring elements is easier, cheaper, faster and operationally more convenient.
- This technique can be used when sampling frame of elements is not available and it is too expensive and time consuming to prepare such a list.
- It is less costly than SRS due to savings of time of journey, identification etc.

Disadvantages :

- In general cluster sampling is less efficient than SRS. However, efficiency of cluster sampling increases with decrease in cluster size and increase in variation within cluster.

Note : Various sampling procedures can be applied to cluster sampling by treating clusters.

10.7.2 Optimum Cluster Size

Efficiency of cluster sampling increases with the increase in the number of clusters. However, the cost of survey also increases with increase in the number of clusters. But in any survey our resources are always limited. That is we must try to minimize cost. Hence it is necessary to determine a balancing point by finding out the optimum cluster size and the number of clusters in the sample so that maximum efficiency can be obtained for a given cost or minimum cost is incurred for a given level of efficiency.



CHECK YOUR PROGRESS

Q 13: State whether the following statements are true or false.

(i) A cluster consists of a number of elimentary units of the population. (T/F)

(ii) Efficiency of cluster sampling increases as cluster size increases. (T/F)

Q 14: In what situation cluster sampling is preferred?



10.8 LET US SUM UP

- Population is a group of individuals with respect to certain characteristic under study.
- Sample is a part of the population selected from it in order to study its characteristic.
- Census or complete enumeration refers to enumeration of each and every unit of the population.
- Sample survey refers to enumeration of each and every unit of the sample.
- Two basic principles of sample survey are validity and optimization.
- Non-sampling errors are present in both census as well as sample survey, whereas sampling errors are present only in sample survey.
- Sampling is broadly classified into two categories– probability sampling and non-probability sampling.
- Stratified random sampling is used when the population under consideration is not homogneous with respect to the characteristic under study.
- stratified random sampling is more efficient than SRS.
- In systematic sampling techneque, only the first unit is selected at random and the rest gets automatically selected according to some pre-designed pattern.
- In systematic sampling when the population size N is a multiple of the sample size n, linear systematic sampling is used and when N is not

a proper multiple of n, circular systematic sampling is to be used to overcome the difficulty of varying sample size.

- In cluster sampling, the population is divided into some non-overlapping clusters and required number of clusters are selected by equal or unequal probabilities of selection.
- Efficiency of cluster sampling increases with the increase in the number of clusters, but at the same time cost also increases. Hence, it is necessary to balance between them so that maximum efficiency can be abtained.

10.9 FURTHER READING

1) Cockran, W. G. Sampling Techniques (1963), John Wilay and Weley Eastern.

2) Goon, A. N, Gupta, M. K, Dasgupta, B. Fundemantals of Statistics (Vol II). (1986). The World Press Private Limited.

3) Singh, Daroga, Chaudhury, F. S. Theory and Analysis of Sample Survey Designs. (1986). New Age International Publishers.

4) Sukhatme, P. V. and Sukhatme, B. V. Sampling Theory of Surveys with Applications (1970). FAO and Asia Publishing House.

10.10 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) Population census, (ii) infinite.

Ans to Q No 2: (i) True, (ii) True, (iii) False.

Ans to Q No 3: Sampling is inevitable under the following conditions.

(i) When the resources for survey, i.e. money and trained man power is limited,

(ii) When there is time constraint,

(iii) When the population is very large,

(iv) When the survey is destructive in nature,

(v) When survey is to be conducted in a very wide area.

Ans to Q No 4: The main objective of the survey is to obtain maximum information about the population at minimum cost, time or labour.

Ans to Q No 5: The difference between parameter and statistic is as follows:

(i) The parameter are population characteristics such as population means, population variance but statistics are sample characteristics such as sample mean, sample variance etc.

(ii) Parameters are constant quantities but statistics are variables.

Ans to Q No 6: (i) True, (ii) False.

Ans to Q No 7: (i)  (ii) 

Ans to Q No 8: If the units of the population are drawn one by one in such a way that the units drawn are returned back to the population after each draw, it is known as SRSWR. In this type of sampling the probability of drawing a unit at each draw remains . In this case a unit may appear more than once in the sample.

On the otherhand if the units selected from the population are drawn one by one but are not returned back to the population, it is called SRSWR. Here the probability of drawing any unit at the r^{th} draw is . In this type of sampling the same unit can never be included in more than one sample.

Ans to Q No 9: (i) False, (ii) True.

Ans to Q No 10: (i) True, (ii) False.

Ans to Q No 11: Systematic sampling is preferred when informations to be collected about trees in a forest, houses in a block, entries in a register, which are in serial order.

Ans to Q No 12: Systematic sampling resembles stratified sampling in the sense that the population is divided into strata of size k where one unit is drawn from each stratum. But the fact that the strata should be homogeneous is not followed here.

Again no independent samples are drawn from each group. Also in systematic sampling we have only one sample from the whole population.

From above discussion it is clear that systematic sampling is quite different from stratified sampling.

Ans to Q No 13: (i) True, (ii) False.

Ans to Q No 14: The cluster sampling is preferred when

(i) the sampling frame is not available or too expensive and time consuming to prepare it,

(ii) the sampling units are distance apart,

(iii) the elimentary units may not be easily identifiable and locatable.

10.11 MODEL QUESTIONS

Q 1: Give a brief account of advantages and disadvantages of sampling.

Q 2: What is simple random sampling? What are its advantages?

Q 3: What is stratified random sampling? Why stratified random sampling is more advantageous than SRS?

Q 4: Distinguish between random sampling and judgement sampling.

Q 5: Distinguish between linear and circular systematic sampling.

Q 6: Describe briefly the different sources of error in sample survey.

Q 7: What are the advantages of sample survey over complete enumeration.

Q 8: What is simple random sampling? Describe two methods of selection of selection of simple random sample.

Q 9: A simple random sample of 30 households was drawn from a city having 14848 households. The number of in the selected households are :

5, 6, 3, 3, 2, 3, 3, 4, 4, 3, 2, 7, 4, 3, 4, 4, 3, 3, 4, 4, 3, 3, 1, 2, 4, 3, 4, 3, 4, 3. Estimate the population total.

Q 10: Write short notes on: (i) Principles of sample survey, (ii) Non-sampling error. (iii) Cluster sampling, (iv) Systematic sampling.

*** ***** ***

UNIT 11: CHARACTERISTICS OF A GOOD ESTIMATOR

UNIT STRUCTURE

11.1 Learning Objectives

11.2 Introduction

11.3 Estimators and Estimates

11.4 Types of Estimation

11.5 Criteria for a Good Estimator

- 11.5.1 Unbiasedness
- 11.5.2 Consistency
- 11.5.3 Efficiency

11.6 Let Us Sum Up

11.7 Further Reading

11.8 Answers To Check Your Progress

11.9 Model Questions

11.1 LEARNING OBJECTIVES

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

- discuss how the theory of estimation is used to draw inference about population characteristies from the information provided by the sample
- distinguish an estimator from an estimate
- explain the difference between point estimation and interval estimation
- select an appropriate statstic to serve as a best estimator of a population parameter.
- have knowledge about the propertices that a good estimator should possess.
- put forward the concept of unbiasedness, consistency and efficiency of an estimator.

11.2 INTRODUCTION

This unit deals with the basic concepts of the theory of estimation. We have already discussed about the terms like population, sample, parameter and statistic etc in unit a of this book. So in the light of these knowledge we may define estimation as the technique of coming to a conclusion, regarding the magnitude of a parameter based on the information provided in the sample. As parameters reflect the characteristis of the population, whose values are unknown, our only way to estimate their values is to depend on the random sample from the same population. Obviously greater the size of the sample better will be the estimate of the parameter.

If we consider the case of a normal distribution with parameters and

of functions of the sample values, called statistic, which we may propose as the estimates of the parameters. But the best estimate will be the one which will be the nearest to the true value of the parameter to be estimated.

In the subsequant sections of the unit, we shall discuss about a few related terms and concepts like estimators and estimates, types of estimation to develope a clear understanding of the topic, Also we are going to learn the concept of unbiasedness, consistency and efficiency with the help of examples.

11.3 ESTIMATORS AND ESTIMATES

In the method of estimation we are interested in finding an estimated formula of the estimate of the parameter, which is called the **'estimator'**. Again the numerical value of the parameter obained on the basis of the random sample using that estimator is called the **'estimate'**.

Thus estimator is the sample statistic used for obtaining an estimate of the parameter being estimated. For example, we know that the sample mean is used to estimate mean , then

estimate refers to the actual sample result used to represent

sample of size n = 36, the avarage age is 20 years, these sample mean

20 years is the estimate.

11.4 TYPES OF ESTIMATION

There are two types of estimation for a population parameter. They are,

(i) Point estimation

and (ii) Interval estimation.

In the first type of estimation, in the light of the sample, one may put forward a single value (i, e a point) as the estimate of the parameter. That is why it is called point estimation.

Sometimes instead of finding a point estimator, one may be interested in finding an interval within which the true value of the parameter lies. The method of construction of such an interval is known as interval estimation. To understand the difference betwen two types of estimation, think of an enumerator collecting data on the number of members per family, i.e, average family size in Assam. On the basis of his sample of say, 1000 familices, the enumerator may find the average family size may be taken as 4.2. This is point estimation.

On the otherhand, again, on the basis of the same sample, he may find that the average family size supposed to lie between 3.9 to 4.6, so the interval 3.9 to 4.6 is the interval estimate.

11.5 CRITERIA FOR A GOOD ESTIMATOR

It follows that an estimator must possess some desirable properties to be a good estimator. Such properties must meet the criterion that a good estimator is one whose sampling distribution is clustered around the true value of the parameter being estimated, than that of any other estimator, so a statistic can be regarded as a good estimator of a parameter if it satisfies the following properties.

(i) Unbiasedness

(ii) Consistency

(iii) Efficiency.

(iv) Sufficiency

We shall discuss about these properties one by one.

11.5.1 Unbiasedness

It is very likely that, the value of a statistic calculated from a given sample will be above or below the actual values of the population parameter of interest, because of the presence of sampling error. So what is desirable is that, the expected value or mean of the possible values of a statistic estimated from over all possible samples should be equal to the population parameter being estimated. If this is true then the statistic is called are unbiased estimator of the population parameter.

So the sample statistic

the population parameter

where, is the mean of the sample statistic

In the sampling distribution of sample mean and sample proportion, we get,

and

Therefore, and

corresponding population parameters and P respectively. Similarly, the difference , i.e. the difference between two sample means, , the difference between two sample proportions are unbiased estimators of population parameters and respectivaly.

Again, if, , then

parameter , and the amount of bias is,

Again, the sample mean square, not the sample variance is an unbiased estimators of the population variance

i.e.,

Where,

square, However, for large n, the sample variance, i.

becomes an unbiased estimator of the population variance

It may be however mentioned that many biased estimators tend to become unbiased estimators asymptotically, because the bias vanishes as n increases.

An important point to be mentioned here is that an unbiased estimator may not be unique. To explain this let us consider the case of a poisson distribution with parameter

As we know for Poisson distribution mean =

Now, as discussed earlier, the sample mean is an unbiased estimator of the population mean.

i.e.

And sample mean square is an unbiased estimator of the population variance.

i.e.

So both sample mean and sample mean square are unbiased estimators of the population parameter



CHECK YOUR PROGRESS

Q 1: State true or false.

(i) The estimating function of a parameter is the estimator. (T/F)

(ii) An estimate is said to be an unbiased estimate of the parameter , if . (T/F)

(iii) In sample from a normal population, the sample mean is an unbiased estimate of the population mean . (T/F)

(iv) In a random sample from a normal population, the sample variance is an unbiased estimate of the population variance. (T/F)

Q 2: Distinguish between estimator and an estimate.

Q 3: What is the difference betwen point estimator and an interval estimator.

11.5.2 Consistency

The property of an estimator being consistent is derived from the law of large numbers which states that, as sample size increases, the sample mean tends to the population parameter ,
a point estimator is consistent, if its value, say,
closer to the population parameter
So we can define a consistent estimator as follows :

Let be an estimate of
then will be a consistent estimator of
as

That is the probability that approaches is 1 as n becomes larger and larger, i.e. higher the sample size, more concentrated should be the distribution be around

For a symmetrical distribution, we know that both sample mean and sample median
population mean
consistency is the only desirable criterion to test for a good estimator.
so, we see that the sample mean
population mean because

Again, the sample proportion
the population proportion P since as

As a matter of fact, it is seen, that consistency is the desirable property of an estimator when the sample size is large. It is of no pratical signifcance when the sample is small.

Remark :

(i) Consistency and unbiasedness of an estimator doesnot necessarily go together, i.e an unbiased estimator need not be consistent and a consistent estimator may be biased.

(ii) Consistent estimators are not unique.

1. If

, show that , is an unbiased estimator of

Solution: since is the sample from

Now,

Now, E(t)

2. Prove that in sampling from a
mean is a consistent estimator of

Solution: We know that, in sampling from

sample mean is also normally distributed with mean

variance

Now,

is consistent estimator of

11.5.3 Efficiency

The efficiency f an estimator is measured in terms of it's variance : for the same population, if there are two consistent estimators for a parameter, then that consistent estimator with

smaller variance is said to be more efficient, because it provides an estimate closer to the population parameter. Thus for a random sample of size n, if and
population parameter , then the relative efficency of
to .

obviously, E < 1.

Where, .

By most efficient estimator, we mean that estimator in a class of consistent estimator for a parameter, whose variance is less than that of any such estimator.

Illustrated Example 1. In a normal sample from a population the sample mean
estimating.

and

Let

then

Efficency of

= 0.64

Again, whenever we deal with unbiased estimators of a parameter, we prefer that estimator which has minimum variance. This leads us to the minimum variance unbiased estimator of

Remark : The correlation co-efficient between a most efficient estimator and any other estimator with efficiency e is

Sufficicncy: An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter. That is sufficiency required that an estimator uses the entire sample informations in estimating the parameter of our interest.

For example the sample statistic to be a sufficient estimator of the population parameter, say,

information and furnish as much of it as possible it obtaining the value of the parameter . A sufficient estimator, if available is always preferred over a non-sufficient estimator, because it allows the use of comprehensive sample information in estimation, so if we say that the sample mean is sufficient estimator of

no other estimator, say x_{med} can convey more information in estimating the parameter

Ex : In a normal population with mean and variance

(i) is a sufficient estimator for , where

(ii) is sufficient estimator for



CHECK YOUR PROGRESS

Q 4: State true or false.

(i) In sampling from a normal population, the sample mean is a consistent estimator of the population mean . (T/F)

(ii) A consistent estimator is unique. (T/F)

(iii) A consistent estimator T of parameter is the most efficient estimator iff V(T) < V ($T^{/}$), where $T^{/}$ is any other consistent estimator of . (T/F)

(iv) If T_1 is the most efficient estimator of a parameter and T_2 be another estimator of , these the efficiency e of T_2 is, (T/F)

Q 5: Prove that, if , where T and $T^{/}$ are both efficient estimators with variance V, then, .

Where p is the correlation co-efficient between T and $T^{/}$.

Q 6: Define consistent estimator. Do consistency and unbiasedness in estimators necessarily go together.



11.6 LET US SUM UP

- Estimation is the technique of coming to a decision about the value of the parameter depending on the imformation provided by the random sample.
- Estimation may be of two types - point estimation and interval estimation.
- The estimating function of a parameter is known as estimator.
- The particular value of an estimator is called the estimate of the parameter.
- The criteria for a good estimator are :
  unbiasedness
  consistency
  efficiency and sufficiancy
- An estimator T is said to be unbiased estimator of the parameter , if, E(T) =
- is a consistent estimator of if converges to
  for large n, where n is the sample size.
- An unbiased estimator need not be consistent and a consistent estimator may be biased.
- The sample mean is an unbiased estimator of the population mean.
- Sample variance is not an unbiased estimator of the population variance of
- If $_1$ and
  other consistent estimator with variauce V_2 then efficiency of
  , e < 1.

- The correlation co-efficient between a most efficient estimator and any other estimator with efficiency e is
- t is called minimum variance unbiased estimor of
  (i) (ii)
  where t' is any other unbiased estimator of

11.7 FURTHER READING

1) Gupta, S. C. and Kapoor, V. K (2002) Fundementals of Mathematical Statistics, Sulten Chand and sons.
2) Goon, A. M, Gupta, M. (1985). An Outline of Statistical Theory, Vol I, New World Press Pvt. Ltd.
3) Mukhopadhyay, P (1996), Mathematical Statistics, New Central Book Agency (P) Ltd.
4) Rao, C. R. (1963), Linear Statistical Inference and its Application, John Willey and sons.

11.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True, (ii) False, (iii) True, (iv) False.

Ans to Q No 2: The estimator is the estimated formula of a parameter. Again, the numerical value of an estimator from a fixed set of values of a random sample is known as estimate. Thus estimate stands for the value of the parameter. For example, the sample mean estimate of population mean ^{2} is an estimate of the population variance , whereas the formula for and s^2 are the estimators.

Ans to Q No 3: The point estimation comprises of finding out an estimated formula of a parameter and finding out the numerical value on the basis of a sample using that estimated formula.

On the otherhand, interval estimation is the method of finding an interval

which covers the true parameter with an assigned probability, say [redacted]. Such an interval is called a confidence interval with confidence co-efficient [redacted]

Ans to Q No 4: (i) True, (ii) False, (iii) True, (iv) False.

Ans to Q No 5: [redacted]

Ans to Q No 6: Let [redacted] be an estimator of [redacted]
then [redacted] is a consistent estimator of [redacted], if [redacted]
[redacted], for. large n, i.e.
[redacted] as [redacted]

Again, consistency and unbiasedness of an estimator doesnot necessarily go together. An unbiased estimator neednot necessarily be consistent. similarly, a consistent estimator may be biased.

11.9 MODEL QUESTIONS

Q 1: Define an estimator and an estimate. State the requirements of a good estimator.

Q 2: Define the following terms and give one example for each,

(i) Consistency,

(ii) Unbiasedness

(iii) Efficiency.

Q 3: Show that the unbiased estimators are not unique with the help of an example.

*** ***** ***

UNIT 12: INDEX NUMBERS

UNIT STRUCTURE

12.1 Learning Objectives

12.2 Introduction

12.3 Index Numbers: Concept and Uses

12.4 Construction of Simple Index Numbers

12.5 Construction of Weighted Index Numbers

- 12.5.1 Laspeyres' Index Number
- 12.5.2 Paasche's Index Number
- 12.5.3 Fisher's Index Number

12.6 Test of Adequacy of Index Numbers

- 12.6.1 Time Reversal Test
- 12.6.2 Factor Reversal Test
- 12.6.3 Circular Test

12.7 Base Shifting, Splicing and Deflating

12.8 Index of Industrial Production (Concept only)

12.9 Let Us Sum Up

12.10 Further Reading

12.11 Answers To Check Your Progress

12.12 Model Questions

12.1 LEARNING OBJECTIVES

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

- describe the uses of index numbers
- discuss the simple and weighted index numbers
- evaluate test of accuracy of index numbers
- discuss the concept of base shifitng, splicing and deflating.

12.2 INTRODUCTION

This unit deliberates on index numbers. Index numbers are the indicators of the various trends in an economy. Price index numbers indicate the position of prices whether they are rising or falling and at what rate.

Similarly, index numbers regarding agricultural production indicates the trend of change whether it is rising or falling at what rate over a period of time. An index number is an economic data figure reflecting price or quantity compared with a standard or base value. The base usually equals 100 and the index number is usually expressed as 100 times the ratio to the base value.

After going through this unit, you will be able to discuss the different concepts relating to index numbers, construct simple and wighted, discuss the concepts of base shifting, splicing and deflating of index numbers.

12.3 INDEX NUMBERS: CONCEPT AND USES

Dear learners, you might have often come across situations like the people making hue and cry over rising prices of goods and commodities. How do we analyse the rising prices in Economics? What tools do we use? To analyse the rising prices, rather change in the price levels, we use the tool 'index numbers' in Economics. Index numbers are one of the most frequently used statistical tools in the field of Economics and other social sciences. Index numbers represents the 'state of happening' in certain variables like the general price level, the wholesale price level, wage rate of labours, retail sales volume, wholesale sales volume, export and imports of a country etc. On such accounts, the index numbers are aptly termed as 'barometers of economic activity'.

We can define an index number as a statistical measure designed to show changes in a variable or a group of related variables with reference to time, geographic location or other entities like income, profession etc. As a statistical measure, an index number is calculated as a ratio of the current value to a base value (also called as reference value) and is expressed in percentage. It should be kept in mind that the index number is more commonly called as an index.

Thus, an index number is a statistical device (measure) with a purpose of showing average changes in one or more related variables (like price or quantity) between two periods of time (say, between 1991 and 1996) or two places like cities (say, Delhi and Mumbai) or countries (say, India and

Japan). For example, we may be interested in knowing which city of India out of Delhi, Mumbai, Kolkata and Chennai is the costliest or the cheapest in terms of price level. A tourist may be interested in knowing about the cost of living at different tourist places. Consumer price index number or cost of living index number helps in taking such decisions.

Let us consider the following table showing the changes in the wholesale prices of primary articles, manufactured products and commodities across all categories in India with reference to the price levels of 1981-82. This means, 1981-82 is the base year prices, i.e., (the index of prices of 1981-82=100).

Table 12.1: Index Numbers of Wholesale Prices in India

Base year 1981-82 (=100)

Period	Primary articles	Manufacture products	All commodities
July, 1992	237.6	222.6	226.6
October, 1992	237.1	229.3	232.4
January, 1993	235.0	228.6	231.6
April, 1993	234.2	231.2	234.6
June, 1993	234.9	233.5	237.7

Source: Centre for Monitoring Indian Economy, July, 1993.

From Table 12.1 it can be seen that compared to the 1981-82 price levels, the price level of primary goods (primary goods comprises food items, non-food items and minerals) in January, 1993 has increased by 234.2 percent or 2.342 times of the 1981-82 price levels of the corresponding period (i.e., January, 1982).

Thus, we have seen that index numbers give us a comparative picture over a period of time. This means that index numbers essentially reflects upon the changes in the variable or the group of related variables over a period of time. For example, if the index of industrial production was 215.1 in 1992-93 (base year: 1980-81), it means that the industrial production in that year increased by 2.151 times compared to 1980-81. It should however be noted that this does not mean that compared to 1980-81, all the sectors of the industry witnessed similar trend of progress. The index of 1992-93 reflects an 'average picture' of the industrial sector. It does not necessary imply that all the sectors of the economy witnessed similar

progress. This means that a few sectors of the industrial sectors might have performed above the average level, while a few might have performed below that level. According to Croxton and Cowden, “Index numbers are devices which measures the change in the level of a phenomenon with respect to time, geographical location or some other characteristics.”

It should be kept in mind that different commodities are measured in different units. For example, rice and wheat are measured in kilograms, cloth in meters and milk in liters etc. Index numbers attempt some averages relating to commodities which are measured in different units. For purpose of comparison, we are interested in knowing relative changes. These relative changes are expressed in percentage terms.

Uses of Index Numbers: Index numbers are constructed with various aims and purposes. Index numbers constructed are put into several uses. Some of them are described below:

- The most commonly used types of index numbers are those that relate to changes in level in prices of a commodity or a group of commodities. Such index numbers are price index numbers. Price index numbers are useful in studying price movements with a view to analyzing their causes, as well as to determine their effects on the economy. To determine economic relationships, it is often useful to compare changes in the market price level with changes in other series, such as gold, bank deposits, bank loans, and the physical volume of production.
- Indices of physical changes over a period of time in production, in marketing, in sales, in imports and exports, and so on have important uses. Such index numbers over a period of time are known as index time series. These index numbers are extremely useful in the study of the movement of the characteristics over time, in the study of seasonal and cyclic behaviour, and in the study of business cycles, etc.
- Index numbers can be used for forecasting future trend in demand of commodities or in volume of production etc. Index numbers of industrial and agricultural production not only reflect the trend but also are useful in forecasting production. Index numbers of unemployment which

reflect the trend of unemployment are useful in determining factors leading to unemployment and in forecasting future trends.

- Index numbers which reflect the general price level, and in particular, the consumer price index numbers are useful in determining the quantum of additional wages or dearness allowances to compensate for the changes in the cost of living. The Government of India and most of the State Governments have index linked formulae for grant of dearness allowances to their employees.
- The changes in consumer price index can be used to determine the real income over a period. Suppose that a person's income in 1980 was 150 times than that in 1970, that is, there is an increase of 50% in money income. Suppose that the cost of living index number in 1980 was 175 times than that in 1970, that is, there is an increase of 75%. Then the person's real income in 1980 is $\frac{150}{175} = 85.7\%$ of his income in 1970. Thus index numbers are useful in determining the real income.



CHECK YOUR PROGRESS

Q 1: State whether the following statements are True (T) or False (F).

(i) Index numbers are specialized averages to measure relative changes. (T/F)

(ii) Index numbers measure net change in related variables over time. (T/F)

(iii) Simple index numbers require weights in their construction. (T/F)

(iv) Index numbers do not indicate the trend of change in the economy. (T/F)

Q 2: Fill in the blanks:

(i) Index number is a statistical device to express ..... change in related variable.

(ii) An index number is a ..... average.

12.4 CONSTRUCTION OF SIMPLE INDEX NUMBERS

In general we are interested in general price level rather than on the price level of any particular item. So we have to consider a number of items at the same time. If each commodity is of different qualities then all such qualities have also to be taken into consideration. We have to consider simultaneously a number of items or commodities together for measuring change in price or production.

The sources of data must be reliable and care should be taken in selecting the sources of data. For example, official publications as well as other reliable unofficial reports from producers and merchants should be considered for computing index for production. If retail food prices are being considered, then prices in stores, not in wholesale markets, are to be taken into consideration.

Now we shall discuss how to construct simple index numbers from a given set of data, like prices of certain commodities, production of certain commodities etc.

There are two methods of constructing simple index numbers. These are:

- Simple aggregate method and
- Method of simple average of relatives.

Remember that all items in a simple index number are assigned equal weights. These are unweighted index numbers.

- **Simple Aggregate Method:** In this case each item is given equal weights. If we give an equal weight to each item it means the same thing, whether each item is given a weight or not. It is the simplest method of constructing index numbers.

  Simple aggregate method to find price index numbers: We use the following three steps to find price index number.

  Step 1: Find the sum of the current year (that is, given period) prices of all the items included in the list. If 1 denote the current year, then find $\sum p_1$.

  Step 2: Find the sum of the base year prices of all the items included in

the list. If 0 denote the base year, then find $\sum p_0$.

Step 3: Use the formula $p_{0/1} = \frac{\sum p_1}{\sum p_0} \times 100.$

Exercise 1: Find the price index number, using simple aggregate method, for the following information.

Commodities	Units	Prices (in Rupees)	
		2000 (year)	2005 (year)
Wheat	Quintal	800	1200
Rice	Quintal	1200	1600
Milk	Litres	12	18
Tea	Kilograms	100	120
Clothing	Metres	30	40
Meat	Kilograms	80	120

Solution: Represent prices of 2005 (current year) as p_1 and that of 2000 (base year) as p_0 and reconstruct the table as follows:

Commodities	p_0	p_1
Wheat	800	1200
Rice	1200	1600
Milk	12	18
Tea	100	120
Clothing	30	40
Meat	80	120
	$\sum p_0 = 2222$	$\sum p_1 = 3098$

Therefore, the price index of 2005 with base year 2000 is:

$$p_{0/1} = \frac{3098}{2222} \times 100 = 139.4$$

Thus the increase in prices (or price level) is 139.40 - 100 = 39.4%.

Exercise 2: The prices of three commodities A, B and C increased from Rs.50, Rs.7 and Rs.3 in 2005 to Rs.52, Rs.8 and Rs.5 respectively in 2006. Using the simple aggregate method, find by how much on an average the prices have increased?

Solution: We know construct the following table:

Commodities	Price in 2005 (p_0)	Price in 2006 (p_1)
A	50	52
B	7	8
C	3	5
	$\sum p_0 = 60$	$\sum p_1 = 65$

Therefore, the price index of 2006 with respect to the base period 2005 is:

$$p_{2005/\ 2006} = \frac{\sum p_1}{\sum p_0} \times 100 = \frac{65}{60} \times 100 = 108.33$$

This shows that the price index of current year 2006 with respect to base year 2005 is 108.33%. Hence the price increase in one year is 108.33 - 100 = 8.33%.

Simple aggregate method to find quantity index numbers: We use the following three steps to find quantity index number.

Step 1: Find the sum of the current year (that is, given period) quantity or volume produced, consumed or transacted of all the items included in the list. If 1 denote the current year, then find $\sum q_1$.

Step 2: Find the sum of the base year quantity or volume produced, consumed or transacted of all the items included in the list. If 0 denote the base year, then find $\sum q_0$.

Step 3: Use the formula $q_{0\ /\ 1} = \frac{\sum q_1}{\sum q_0} \times 100.$

- **Method of Simple Average of Relatives:** In this method, average of the relatives is obtained by using any one of the measures of central tendency. In particular, here we shall use arithmetic mean for averaging the relatives. We shall discuss briefly this method only for price relatives. Similar formulas on quantity index number and value index number can be obtained considering quantity and value relatives instead of price relative. There are few questions to find quantity and value index numbers in the section "Possible Questions". You should work out those problems with the help of the problems on price index numbers which are solved below.

Method of Simple Average of Price Relatives:

We know that a price relative is nothing but the ratio of current year prices to those in the base year. Taking the arithmetic mean of the price relatives, we get the following formula for price index number:

$$p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{N}$$

where N stands for number of commodities included in the index numbers. Steps to calculate price index numbers by Simple Average of Price Relatives Method:

Step 1: Find percentage price relative for each commodity, that is, $\frac{p_1}{p_0} \times 100$

Step 2: Find the sum of these percentage price relatives, that is, $\sum\left(\frac{p_1}{p_0} \times 100\right)$.

Step 3: Divide $\sum\left(\frac{p_1}{p_0} \times 100\right)$ by the total number of commodities included in the list.

Exercise 3: Give the solution of Exercise 1 using the Simple Average of Price Relatives.

Solution: We construct the following table:

Sl No.	Commo-dities	Units	Prices (in Rupees)		Price Relative $\frac{p_1}{p_0} \times 100$
			2000 p_0	2005 p_1	
1	Wheat	Quintal	800	1200	$\frac{1200}{800} \times 100 = 150.00$
2	Rice	Quintal	1200	1600	$\frac{1600}{1200} \times 100 = 133.33$
3	Milk	Litres	12	18	$\frac{18}{12} \times 100 = 150.00$
4	Tea	Kilograms	100	120	$\frac{120}{100} \times 100 = 120.00$

5	Clothing	Metres	30	40	$\frac{40}{30}\times 100 = 133.33$
6	Meat	Kilograms	80	120	$\frac{120}{80}\times 100 = 150.00$
N = 6					$\sum\left(\frac{p_1}{p_0}\times 100\right) = 836.66$

Hence, the price index number of 2005 with respect to base year 2000 is:

$$p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0}\times 100\right)}{N} = \frac{836.66}{6} = 139.44.$$

Exercise 4: Give the solution of Exercise 2 using the Simple Average of Price Relatives.

Solution: We first prepare the following table:

Sl No.	Commo-dities	Prices (in Rupees)		Price Relative
		2005 p_0	2006 p_1	
1	A	50	52	
2	B	7	8	
3	C	3	5	
N=3				

Hence, the price index number of 2006 with respect to base year 2005 is:

$$p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0}\times 100\right)}{N} = \frac{384.95}{3} = 128.32.$$

12.5 CONSTRUCTION OF WEIGHTED INDEX NUMBERS

In weighted index number each item is given weight according to the importance it occupies in the list. There are two groups of methods to calculate index number of this category: (a) Weighted aggregate method and (b) Weighted average of price relative methods. In this unit, we will study three types of weighted aggregate method. These methods are popularly known as (1) Laspeyre's Method, (2) Paasche's Method and (3) Fisher's Method. All the methods are used to calculate weighted index numbers. The main difference between the Laspeyre's Method and Paasche's Method is that Laspeyre's uses base year quantities of commodities as their relative weights, while Paasche's uses current year quantities of commodities as their relative weights for preparing a price index.

12.5.1 Laspeyre's Index Number

Laspeyre's Price Index Number: It uses base year quantities as the weights. Accordingly, the formula is

$$P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Laspeyre's Quantity Index Number: If we multiply the quantities by the base year price, then we get Laspeyre's quantity index number. Accordingly the formula for Laspeyre's quantity index number is:

$$Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

Steps to Calculate Weighted Index Number by Laspeyre's Method:

Step 1: Multiply current year price (p_1) with the base year quantity (q_0) to get $p_1 q_0$ for each item/ commodity.

Step 2: Multiply base year price (p_0) with the base year quantity to get $p_0 q_0$ for each item/ commodity.

Step 3: Add all $p_1 q_0$ and $p_0 q_0$ separately to get $\sum p_1 q_0$ and $\sum p_0 q_0$.

Step 4: Divide $\sum p_1 q_0$ by $\sum p_0 q_0$ and multiply by 100 to obtain Laspeyre's price index number.

Exercise 5: Find Laspeyre's price index number from the following data.

Commodity	Price (in Rs)		Quantity bought in units	
	1995	1996	1995	1996
A	50	52	10	12
B	7	9	3	4
C	3	5	7	6

Solution: Represent prices and quantities in current year by p_1 and q_1 respectively and represent prices and quantities in base year by p_0 and q_0 respectively. Let us construct the following table following the steps 1-3.

Commodities	p_0	q_0	p_1	p_0q_0	p_1q_0
A	50	10	52	500	520
B	7	3	9	21	27
C	3	7	5	21	35
				$\sum p_0q_0$ = 542	$\sum p_1q_0$ = 582

Therefore the Laspeyre's price index number is:

$$P_L = \frac{582}{542} \times 100 = 107.38$$

12.5.2 Paasche's Index Number

Paasche's Price Index Number uses current year quantities *(q_1)* as the weights. Accordingly, the formula is

$$P_P = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100$$

Paasche's Quantity Index Number: It uses current year prices *(p_1)* as the weights. Accordingly, the formula is

$$Q_P = \frac{\sum q_1p_1}{\sum q_0p_1} \times 100$$

Exercise 6: Find Paasche's price index number from the data given in Exercise 5.

Solution: Represent prices and quantities in current year by p_1 and q_1 respectively and represent prices and quantities in base year by p_0 and q_0 respectively. Let us construct the following table.

Commodities	p_0	q_1	p_1	p_0q_1	p_1q_1
A	50	10	52	500	520
B	7	3	9	21	27
C	3	7	5	21	35
				$\sum p_0q_1 = 646$	$\sum p_1q_1 = 690$

Therefore the Paasche's price index number is:

$$P_P = \frac{690}{646} \times 100 = 106.8$$

Remark: You have noticed that the value of Laspeyre's index number (107.38) and the value of Paasche's index number (106.8) are different for the same data. The difference has arisen because of differences in weights.

12.5.3 Fisher's Index Number

The Fisher's price index number is nothing but the geometric mean of the Laspeyre's and Paasche's price index numbers. Hence the formula for Fisher's price index number is:

$$P_F = \sqrt{P_L \times P_P} = \sqrt{\left(\frac{\sum p_1q_0}{\sum p_0q_0} \times 100\right) \times \left(\frac{\sum p_1q_1}{\sum p_0q_1} \times 100\right)}$$

$$= \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100 \qquad \cdots(1)$$

Similarly, the Fisher's quantity index number is given by:

$$Q_F = \sqrt{Q_L \times Q_P} = \sqrt{\left(\frac{\sum q_1p_0}{\sum q_0p_0} \times 100\right) \times \left(\frac{\sum q_1p_1}{\sum q_0p_1} \times 100\right)}$$

$$= \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}} \times 100 \qquad \cdots(2)$$

Exercise 7: Find Fisher's price index number from the data given in Exercise 5.

Solution: We have already found that for the given data,

$P_L = 107.38$ and $P_P = 106.8$.

Hence the Fisher's price index number is:

$$P_F = \sqrt{107.38 \times 106.8} = 107.09$$



CHECK YOUR PROGRESS

Q 3: State whether the following statements are true (T) or False (F).

(i) There are two types of index numbers: simple or unweighted and weighted index numbers.(T / F)

(ii) There are two methods of constructing simple index numbers: simple aggregate method and simple average relatives. (T / F)

(iii) We the same index number whether we use simple aggregate method or simple average of price relatives' method. (T / F)

(iv) Laspeyre's uses current year quantities as weights. (T / F)

(v) Paasche's uses base year quantities as weights. (T / F)

Q 4: Fill up the blanks.

(i) In Laspeyre's method of finding price index number we use ..... quantities as weights.

(ii) In Paasche's method of finding price index number we use ..... quantities as weights.

(iii) $I = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$ is ..... weighted index number.

(iv) $I = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$ is ..... weighted index number.

12.6 TEST OF ADEQUACY OF INDEX NUMBERS

The index numbers as an aggregative relative should satisfy a few properties. We shall examine the properties that a good index number should have.

12.6.1 Time Reversal Test

If the two periods, the base period and the reference period are interchanged, the product of the two index numbers should be unity. In other words, one should be the reciprocal of the other.

Denote the index number of the given year (1) relative to the base year (0) by $I_{0/1}$. Then $I_{1/0}$ is the index number of the year (0) relative to the year (1), that is, with the years interchanged time reversal test requires that

$$I_{0/1} \times I_{1/0} = 1, \qquad or \qquad I_{1/0} = \frac{1}{I_{0/1}}$$

An index number which satisfies the above property is said to satisfy time reversal test.

Remark: As we have seen in section 19.5, all the relatives satisfy the time reversal property.

Let us now examine whether Laspeyre's index number satisfy this property or not. We know that Laspeyre's index number (without percentage) is $\frac{\sum p_1 q_0}{\sum p_0 q_0}$. Interchanging the time subscripts, we get $\frac{\sum p_0 q_1}{\sum p_1 q_1}$ and hence their product is $\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1$.

Consequently, Laspeyre's price index number does not satisfy the time reversal property. Similarly, it can be seen that Laspeyre's quantity index number also does not satisfy the time reversal property. Also, Paasche's price and quantity index numbers do not satisfy this property.

Now consider the Fisher's price index number. Interchanging the time subscripts we find another index number. The product of the two is given by

$$\sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{p_1 q_0}}$$

which is equal to 1. Hence Fisher's price index number satisfies the time reversal property.

12.6.2 Factor Reversal Test

If the two factors *p* and *q* in a price index formula are interchanged, so that a quantity index number is obtained, then the product of the two index numbers should give the true value ratio

$$\frac{\sum p_1 q_1}{\sum p_0 q_0}$$

In other words, the price index number multiplied by the corresponding quantity index number should give the true ratio of value in the given year (1) to the value in the base year (0). This property holds for a single commodity, since

$$\frac{p_1}{p_0} \times \frac{q_1}{q_0} = \frac{p_1 q_1}{p_0 q_0}$$

but it does not hold for most of the index numbers. For example,

$$P_L \times Q_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Now consider the Fisher's index number. We have from the formulas (1) and (2),

$$P_F \times Q_F = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

This proves that Fisher's index number also satisfy factor reversal test. Fisher's index number is one of the rare index numbers which satisfy both the time reversal and factor reversal tests. So it is called "Fisher's Ideal Index Number".

12.6.3 Circular Test

This test is based on the shifting of the base period. If $I_{i/j}$ denotes the index number for the given period *j* with respect to the base period *i* then this test requires that

$$I_{a/b} \times I_{b/c} \times I_{c/a} = 1$$

Remark: As we have seen in section 19.5, all the relatives satisfy the circular property.

Remark: Verify that none of the weighted index numbers satisfy circular test.

12.7 BASE SHIFTING, SPLICING AND DEFLATING

In this section, we shall briefly discuss the concepts of base shifting, splicing and deflating.

Base Shifting: In Statistics, we often require to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the eniture series. The change in reference base period is usually referred to as '*base shifting*'.



ACTIVITY 12.1

The Ministry of Statistics, Government of India has changed the base for the calculation of GDP from time to time. Undertake an exercise on tracking such changes since 1990.

Base shifting is undertaken for two important reasons. a) The previous base (i.e., the base year currently in practice) has become too old is not much relevant for making comparision in the present times. By shifting the series, the series is presented to be 'current' and more relevant to the present times. b) When different data are presented for comparision pertaining to the different time periods, it becomes necessary to first make them comparable with the 'time' dimension. For example, it would be

irrevalent to compare the GDP estimates of a country of 1850, 1920, 1950 and 2010 without any statistical treatment. For making the comparision effective, the data must be first worked out to one base year (say, 2010). This is particularly important when we compare them in one single graph.

Splicing: Splicing means combining two or more series of overlapping index numbers to obtain a single index number on a common base. Sometimes, a series of index numbers is available for a period of time, which has been substantially revised including shifting of the base year. It may also happen that the weights assigned to the series have become out dated. As such, new weights may be assigned and the series is worked out anew. In such cases, two series would result, the old one and the new one. Now, it may become necessary to convert the two series into one continuous series. This is done with the help of splicing.

Splicing can be done only if the index numbers are constructed with the same items and have an overlapping year. Splicing like base shifting will give accurate result of the index numbers, use simple geometric mean and satisfy the circular test. In splicing of index numbers, we find a common factor by which the spliced index number series is multiplied to get a common base.

Deflating: Deflating means adjusting, correcting or reducing a value which is inflated. Thus, in defilating, we make adjustments for the effect of changing price levels. You know that a rise in price levels essentially reduces the real purchasing power of money. So, to compare the value of money (or its purchasing power), in different time peirods, we need to adjust the rate of inflation during the given peirod. This is done through deflating. Deflating is a technique of converting a series of values calculated at current prices into constant prices of a given year. This is particularly desirable in the case of an economy which has inflationary trends because in such an economy, the increase in the prices of commodities over a period of years means a fall in their real incomes. For example, a rise in money income may not amount to a rise in real income. As such, it becomes necessary to adjust or correct nominal wages in accordance with the rise in the corresponding price index to arrive at the real income. The purchasing power is given by

the reciprocal of the index number and consequently the real income is obtained on dividing the money or nominal income by the corresponding appropriate price index and multiplying the result by 100.

12.8 INDEX OF INDUSTRIAL PRODUCTION (CONCEPT ONLY)

The Index of Industrial Production (IIP) is an index which shows the growth rates in different industry groups of the economy in a stipulated period of time. The IIP index is computed and published by the Central Statistical Organisation (CSO) on a monthly basis.

Description: IIP is a composite indicator that measures the growth rate of industry groups classified under,

(1) Broad sectors, namely, Mining, Manufacturing and Electricity

(2) Use-based sectors, namely Basic Goods, Capital Goods and Intermediate Goods.

Currently IIP figures are calculated considering 2004-05 as base year.



12.9 LET US SUM UP

- Index number is a statistical measure or device with a purpose of showing average change in one or more related variables over time and space.
- Price index numbers are more commonly used. It measures relative changes in prices over a time period.
- We can have either simple or weighted index number. Simple index is also called unweighted index number or index number with equal weights. In this unit, we have discussed about the simple index numbers as well as weighted index number.
- We have learnt about two methods of constructing simple index numbers: Simple aggregate method and Method of simple average of relatives. The formula to find index number using simple aggregate

method is given by $p_{0/1} = \frac{\sum p_1}{\sum p_0} \times 100.$

Again, the formula to find price index number using the method of simple averages of relatives is given by $p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{N}.$

- We have also leant about weighted index numbers. There are two groups of methods to calculate index number of this category: (a) Weighted aggregate method and (b) Weighted average of price relative methods. We briefly discussed about three types of weighted aggregate method. These methods are popularly known as (1) Laspeyre's Method, (2) Paasche's Method and (3) Fisher's Method.
- We have also learnt about the test of adequacy of index numbers. We mentioned three properties that a good index number should have. Fisher's index number satisfies both the time reversal and factor reversal tests and hence it is known as Fisher's Ideal Index Number.

12.9 FURTHER READING

1. J. Medhi, *Statistical Methods, An Introductory Text*, New Age International (P) Limited, Publishers, 2nd edition, 2000.

12.10 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True (ii) True (iii) False (iv) False

Ans to Q No 2: (i) average (ii) specialized

Ans to Q No 3: (i) True (ii) True (iii) False (iv) False (v) False

Ans to Q No 4: (i) base year (ii) current year (iii) Laspeyre's

12.11 MODEL QUESTIONS

Q 1: Explain the meaning of an index number and its main characteristics.

Q 2: "Index numbers are economic barometers". Why are index numbers so called? Explain.

Q 3: Explain the method of construction of unweighted index numbers by Simple Aggregate Method.

Q 4: Explain the method of construction of unweighted index numbers by Simple Aggregate of Price Relative Method.

Q 5: The exchange rate of a US dollar was Rs.44.00 in July 2008 and was Rs.50.00 in February 2009. Find the price relative.

Q 6: The production of rice in Assam in the month of January 2008 was 3608 tonnes and that in February 2009 was 3700 tonnes. Find the quantity relative.

Q 7: The following table shows the prices of the commodities consumed in 1970 and 1978.

Commodities	Unit	1970 Price (Rs)	1978 Price (Rs)
A	kilogram	4.00	4.60
B	litre	8.00	8.75
C	metre	7.00	8.00
D	kilogram	14.00	16.00
E	litre	3.00	4.00

Find the price index number using:

(a) simple aggregate method

(b) method of simple aggregate of price relatives.

Q 8: The production of rice, milk, cloth, wheat and petrol are given in the following table. Taking 1980 as base year, compute quantity index number using:

(a) simple aggregate method

(b) method of simple aggregate of quantity relatives.

Commodities	Unit	Quantity produced	
		1980	1990
Rice	kilogram	400000	700000
Milk	litre	800000	905000
Cloth	metre	100000	105000
Wheat	kilogram	140000	160000
Petrol	litre	300500	400000

Q 9: What do you mean by tests of index numbers?

Q 10: Explain the important criterion or test that a good index number should satisfy.

Q 11: The prices and quantity of five commodities are given in the following table. Taking 2000 as base year, compute value index number using:

(a) simple aggregate method

(b) method of simple aggregate of value relatives.

Comm-	2000		2002	
odities	Price (Rs)	Quantity (Kg)	Price (Rs)	Quantity (Kg)
A	400	10	700	12
B	800	15	1905	17
C	100	5	125	5
D	140	7	160	7
E	300	6	400	6

Q 12: Find Laspeyre's and Paasche's price and quantity index number from the following data. Also find the Fisher's price and quantity index numbers.

Commo-	Base year		Current year	
dities	Price (Rs) p_0	Quantity (kg) q_0	Price (Rs) p_1	Quantity (kg) q_1
A	5	25	6	30
B	10	5	15	4
C	3	40	2	50
D	6	30	8	35

Q 13: The following table shows the price quantity of the commodities consumed in 2000 and 2006.

Commo-dities	2000		2006	
	Price (Rs)	Quantity (kg)	Price (Rs)	Quantity (kg)
I	400	35	460	35
II	800	10	1000	10
III	300	15	50	12
IV	4000	10	5000	8
V	50	20	300	20

Find the following index numbers.

(1) Laspeyre's Price and Quantity index numbers.

(2) Paasche's Price and Quantity index numbers.

(3) Fisher's Price and Quantity index numbers.

*** ***** ***

UNIT 13: TESTING OF HYPOTHESIS

UNIT STRUCTURE

13.1 Learning Objectives

13.2 Introduction

13.3 Statistical Hypothesis: Null Hypothesis and Alternative Hypothesis

13.4 Errors in Hypothesis Testing, Level of Significance and Critical Region

13.5 One -Tailed and Two-Tailed Tests

13.6 General Procedure for Hypothesis Testing

13.7 Testing of Hypothesis in case of Large Samples

- 13.7.1 Hypothesis Testing for Single Population Mean
- 13.7.2 Hypothesis Testing for Single Population Proportion
- 13.7.3 Hypothesis Testing for Difference between Two Population Means
- 13.7.4 Hypothesis Testing for Difference between Two Population Means

13.8 Let Us Sum Up

13.9 Further Reading

13.10 Answers To Check Your Progress

13.11 Model Questions

13.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- establish null and alternative hypotheses about a population parameter
- understand the two types of errors; interpret the confidence level and the significance level
- develop hypothesis testing procedures for accepting or rejecting a null hypothesis
- discuss how and when to use the normal and t distributions for hypothesis testing

13.2 INTRODUCTION

In today's dynamic business environment, decision makers need might answers to certain questions in order to take optimum decisions. For example a pharmaceutical concern might be interested in assessing whether a new drug is really effective in curing a disease; a personnel manager might be interested to study the job satisfaction level of employees, a marketing manager might be interested in knowing the customer satisfaction about a particular product, etc. In all the situations, the concerned manager has to take wise decisions since such decisions involve the risk of taking wrong decisions. It is here that modern theory of probability plays a vital role in decision making and the branch of Statistics that develops the criterion for such decisions is known as testing of hypothesis. Hypothesis testing begins with an assumption commonly termed as "hypothesis" that a researcher makes about an unknown population parameter. Hypothesis testing is a well-defined procedure which enables us to decide objectively whether to accept or reject the hypothesis based on the information obtained from the sample.

13.3 STATISTICAL HYPOTHESIS: NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

A statistical hypothesis is some assumption or statement, which may or may not be true, about an unknown population parameter. For example, suppose the Vice President (HR) of a company wants to examine the effectiveness of a training programme which the company has organized for all its employees located at different places of the country. The Vice President (HR) will then make an assumption that "training has not enhanced efficiency". To test such claim statistically, sample data are collected and analyzed. The assumption is accepted or rejected on the basis of the sample findings.

There are two types of statistical hypothesis: (i) Null hypothesis and (ii) Alternative hypothesis. In hypothesis testing, to reach decisions we must start with a hypothesis, called null hypothesis which is symbolized H_0. The

null hypothesis asserts that there exists no (significant) difference between the sample statistic and the population parameter and whatever observed difference is there is merely due to fluctuations of sampling from the same population. In the words of R.A. Fisher, "Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true." The null hypothesis that makes a claim regarding the specific value of the parameter is generally expressed by any one of the following mathematical statements

(i) $H_0 : \mu = \mu_0$ (ii) $H_0 : \mu \leq \mu_0$ (iii) $H_0 : \mu \geq \mu_0$

where μ is the population mean and μ_0 denotes a hypothesized value of μ.

An alternative hypothesis, generally denoted by H_1, is the opposite of the null hypothesis. Thus any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. It is obvious that both null and alternative hypotheses cannot be true simultaneously at the same time. In other words, when null hypothesis is found to be true, the alternative hypothesis is to be rejected or when null hypothesis is rejected, the alternative hypothesis must be true. For example, if we are interested in studying the hypothesis that the population has a specified mean, μ_0 (say), then we have the null hypothesis $H_0 : \mu = \mu_0$ and against this null hypothesis we may have any one of the following alternative hypotheses (depending upon the situation) given by,

(i) $H_1 : \mu \neq \mu_0 \quad (i.e., \mu > \mu_0 \text{ or } \mu < \mu_0)$

(ii) $H_1 : \mu > \mu_0$ (iii) $H_1 : \mu < \mu_0$

The alternative hypothesis given by (i) is called two-tailed alternative and the alternatives given by (ii) and (iii) are called right-tailed and left-tailed alternatives respectively. Accordingly, the nature of the alternative hypothesis enables us to decide whether we have to use an one-tailed (right or left) or two-tailed test.

Simple and Composite hypothesis: If the hypothesis completely specifies the population, then it is called simple hypothesis, otherwise it is called composite hypothesis. For example, from a normal population $N(\mu, \sigma^2)$, the hypothesis of the form

$H: \mu = \mu_0$ and $\sigma^2 = \sigma_0^2$ (both μ and σ^2 are specified)

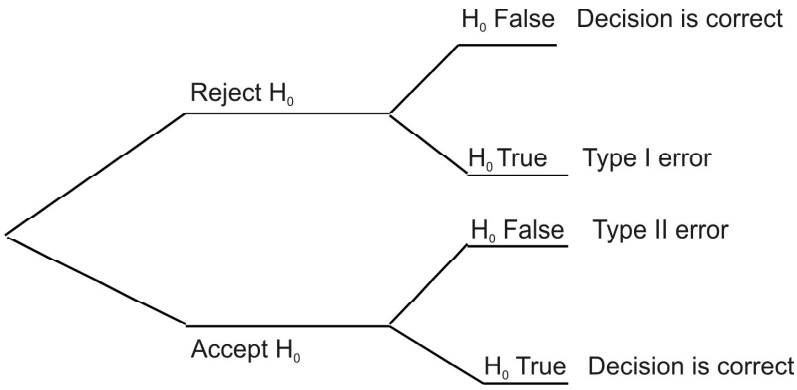
is a simple hypothesis because it completely specifies the distribution. On the other hand, the hypotheses of the forms

(i) $H: \mu = \mu_0$ (σ^2 is not specified), (ii) $H: \sigma^2 = \sigma_0^2$ (μ not being specified) are composite since each of these does not completely specify the value of the parameter.

13.4 ERRORS IN HYPOTHESIS TESTING, LEVEL OF SIGNIFICANCE AND CRITICAL REGION

In hypothesis testing, the null hypothesis H_0 should be accepted if it happens to be true and the null hypothesis H_0 should be rejected if it happens to be false. However, in practical situations, the correct decision is not always possible while testing a null hypothesis. Since the decision to reject or accept a null hypothesis depends only on sample data therefore a decision maker may commit two types of errors (i.e., take wrong decisions) which are diagrammatically shown in Fig.13.1.

Figure 13.1: Type I and Type II errors



Type I error is committed if we reject a null hypothesis when it is true. The possibility of committing Type I error is denoted by the symbol α. Thus,

$\alpha = P$ (Type I error)= P (Reject H_0 when H_0 is true)

On the other hand, Type II error is committed if we accept a null hypothesis when it is false. The possibility of committing Type II error is denoted by the symbol β. Thus we have

$\beta = P$ (Type II error)=(Accept H_0 when H_0 is false)

where α and β are also known as the sizes of Type I and Type II error respectively.

For example, a business firm, before introducing a new product, has to choose one of the two decisions- to introduce or not to introduce the product. Again if the product is introduced then the firm may face two situations- the failure of the product and the success of the product. The firm thus runs the risk of taking wrong decisions in two ways. The one risk of taking wrong decision known as Type II error, is that the firm does not introduce the product though it would have succeeded had it been introduced. The another risk of taking wrong decision known as Type I error is that the firm introduces the product but it failed.

In the terminology of Statistical Quality Control, Type I error is known as producer's risk while Type II error is known as consumer's risk.

In practice, it is not possible to make error free judgment in Statistics. Although both the errors should be very small yet it is not possible to minimize both these errors simultaneously at the same time on the same hypothesis test. Generally there exists an inverse relationship between α and β as a result of which an attempt to reduce type I error (α) leads to an increase in Type II error (β) and vice-versa. Since in most of decision making problems, it is more risky to accept a wrong hypothesis than to reject a correct one, i.e., consequences of Type II error are likely to be more serious than the consequences of Type I error therefore a compromise is made by minimizing the more serious error (i.e., β) after keeping fixed the less serious error (i.e., α).

In testing of hypothesis, the **Level of significance** is defined in terms of the probability of rejecting a null hypothesis when it is true. The level of significance is also known as the size of the critical region or the rejection region and generally denoted by α . It is to be noted that the level of significance should be determined before a test procedure so that obtained result is free from the influence of the decision maker. In practice, generally adopted level of significance by researchers is: 1 % (=0.01), 5% (=0.05) and 10% (=0.10). If a researcher use 5% level of significance, it would

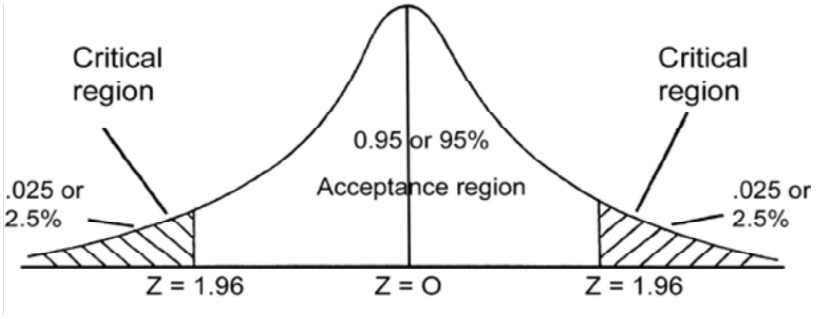
mean that in 5 samples out of 100 he is likely to reject a correct null hypothesis H_0. In other words the researcher is 95% confident that his decision to reject H_0 is correct. Thus higher the significance level adopted for testing a hypothesis, the higher the chance of rejecting a true null hypothesis.

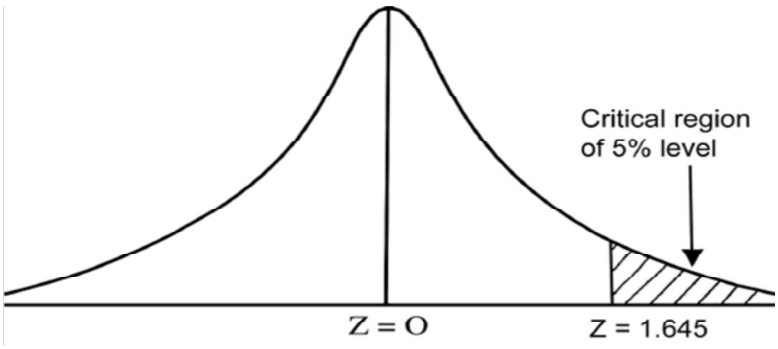
In hypothesis testing, the area under the sampling distribution curve of the test statistic is divided into two mutually exclusive regions termed as acceptance region and rejection or critical region. The acceptance region is a range of values of the sample statistic that lead us to the acceptance of null hypothesis H_0 while critical region is the range of sample statistic values within which if values of the sample statistic falls lead to the rejection of null hypothesis H_0. In Figure 13.2 if the calculated value of the sample statistic falls in the acceptance region, the null hypothesis is accepted otherwise it is rejected. For making a decision whether to accept or reject a null hypothesis, a researcher has to determine the critical value that separates the regions of acceptance and rejection. The determination of critical value depends on the size of the rejection region, which is directly associated with the risk involved in decision making.

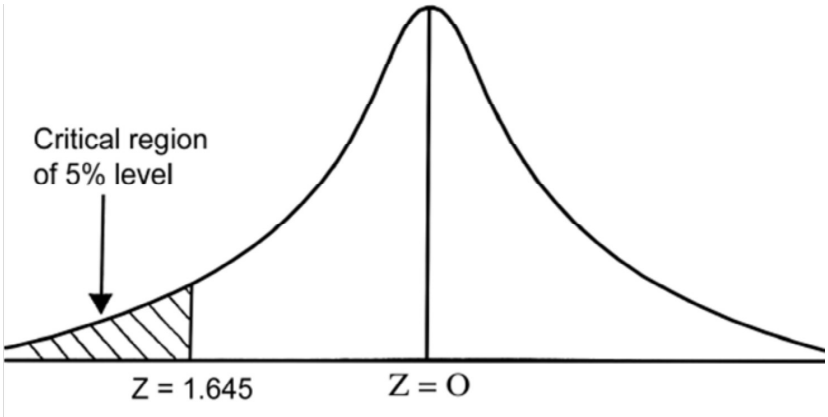
13.5 ONE TAILED AND TWO TAILED TESTS

In any test, there are two types of tests of hypothesis. They are: (i) One-tailed tests and (ii) Two-tailed tests. Whether to use a one-tailed test or two tailed test depends entirely on the formulation of the alternative hypothesis. If the alternative hypothesis is one-tailed then the test to be applied is one-tailed and if the alternative hypothesis is two-tailed then the corresponding test is two-tailed. In case of two-tailed test, the critical or significant values of the test statistic lie towards both the tails of the graph of the sampling distribution of the statistic. Figure 13.2 relates to standard normal test statistic Z. The figure reveals that critical values of the standard normal test statistic Z. lies on both sides of the mean i.e., in both the tails of the distribution. The shaded region under the normal curve is the rejection region (or critical region) at 5% level of significance, the region being 5% of the total area. In other words, the total area under the normal curve being unity, the size of the rejection region is 0.05.

Again sometimes we may encounter certain situations where alternative hypothesis states that either one process is better than the other or one process is worse than the other. But in two-tailed test the alternative hypothesis H_1 states that both may occur i.e., one process may be better or worse than the other. Figure 13.3 and Figure 13.4 below exhibits the critical values and critical regions of the standard normal test statistic for right-tailed test and left-tailed test respectively for 5% level of significance.







Now, we present below a summary of certain critical values of test statistic Z (i.e., standard normal test statistic) for both one-tailed and two-tailed tests at some commonly used level of significance. These critical values of Z for different levels of significance have been obtained by using the table of area under normal curve.

Table 13.1: Certain Critical Values of Test Statistic Z

Critical values (Z_α)	Level of significance (α)		
	5% or 0.05	1% or 0.01	10% or 0.010
Two-tailed test	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.645$
Right-tailed test	$Z_\alpha = 1.645$	$Z_\alpha = 2.33$	$Z_\alpha = 1.28$
Left-tailed test	$Z_\alpha = -1.645$	$Z_\alpha = -2.33$	$Z_\alpha = -1.28$

Note: (i) The notation Z_α stands for the critical value of Z at α level of significance.

(ii) Critical or significant values are also known as tabulated values.

(iii) If the size of the sample is small (usually less than 30) then the normality assumption of the sampling distribution of the test statistic is not valid and hence we cannot use the significant values given in the above table. In such situation we use the significant values based on the exact sampling distribution of Z, which turns out to be t, F or χ^2. (Please refer to any Standard textbook on Statistics).

13.6 GENERAL PROCEDURE FOR HYPOTHESIS TESTING

Hypothesis testing is a well-defined systematic procedure that involves five different steps as given below:

(1) The first step is to set up the null hypothesis H_0.

(2) Set up the alternative hypothesis H_1. The nature of alternative hypothesis H_1 helps us to choose whether to be adopted one-tailed or two-tailed test.

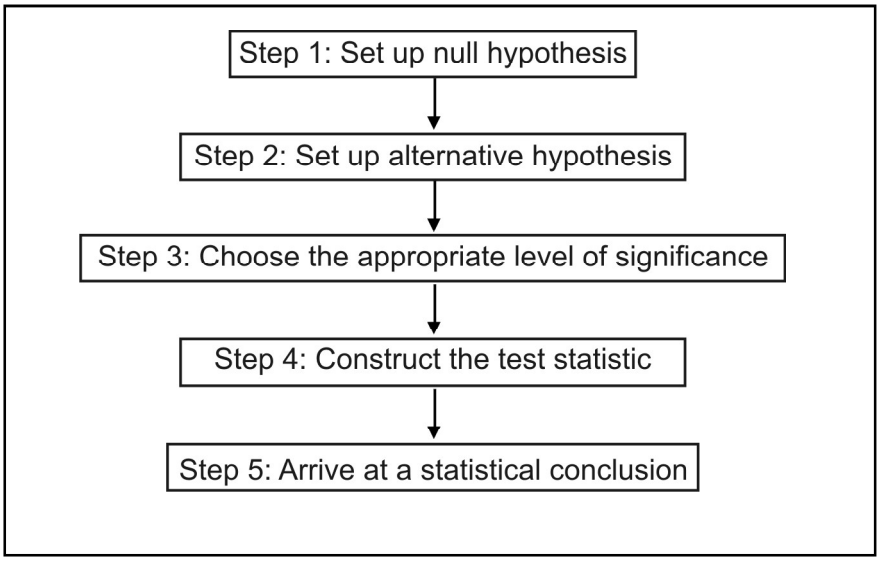
(3) To choose an appropriate level of significance. Commonly used levels of significance are: 0.01, 0.05 and 0.10.

(4) After setting up the null and alternative hypotheses, the researcher needs to select an appropriate statistical test that will be used for analyzing statistical data. Type, number and the level of data may provide a platform in the selection of the statistical test. Apart from these, the statistics used in the study (mean, proportion and variance etc.) must also be considered when a researcher selects an appropriate test, which can be used for testing of hypothesis to obtain the best results. Some of the widely used testing procedures are: z, t, F and χ^2.

(5) The final step in any hypothesis testing is to draw a statistical decision, which involves either to accept or to reject a null hypothesis. This depends on whether the calculated value of the test statistic falls in the acceptance region or in the rejection region at a given level of significance. In other words, if the computed value of Z (in modulus value) is less than the critical value Z_α then we say that it is not significant and null hypothesis may be accepted. On the other hand, if the computed value of Z (in modulus value) is greater than the critical value Z_α then we say that it is significant and null hypothesis is rejected at the level of significance α.

Therefore, various steps to be followed in testing of a statistical hypothesis can be summarized as shown in Figure 13.5

Figure 13.5: Steps in testing of a Statistical Hypothesis



Power of Test: Another way to evaluate the goodness of a statistical test is the complement of type II error, known as power of a test, which is defined as:

Power of a test = P (rejecting H_0 when it is false)

= P (rejecting H_0 when H_1 is true)

= 1- P (accepting H_0 when H_1 is true)

=1- P (type II error)

=1- β

Where β is the probability of type II error. Among a class of tests, the best is the one which has the maximum power for the same size.



CHECK YOUR PROGRESS

Q 1: What is meant by Hypothesis testing?

.....

.....

Q 2: What is neant by Type I error?

.....

.....

Q 3: What is meant by Type II error?

.....

.....

13.7 TESTING OF HYPOTHESIS IN CASE OF LARGE SAMPLES

Here we shall develop the tests of significance in case of large samples. In practice, a sample is to be considered as large if the sample size is greater than or equal to 30 i.e., $n \geq 30$. Now since the assumptions made in case of large samples are not valid for small samples therefore the tests of significance used to deal with problems relating to large sample are different from the ones used for small samples.

The following are the assumptions made in testing of hypothesis relating to large samples:

(i) The population from which the sample is drawn is normally distributed. Consequently the sampling distribution of the sample statistic is also

normal.

(ii) Since sample size is large, due to central limit theorem, the value given by any random sample can be used in place of population value for calculating the standard error of the estimate.

13.7.1 Hypothesis Testing for Single Population Mean

Suppose a random sample of size $n(n \geq 30)$ with mean $\bar{x}$ is drawn from a population. We hold the claim that the population mean μ has the specified value μ_0.

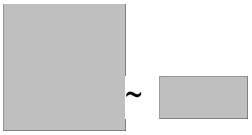
Here we set up the null hypothesis $H_0 : \mu = \mu_0$. Against this hypothesis we may have any one of the following three alternative hypotheses

(i) $H_1 : \mu \neq \mu_0$ (Two-tailed test)

(ii) $H_1 : \mu > \mu_0$ (Right-tailed test)

iii) $H_1 : \mu < \mu_0$ (Left-tailed test)

Under the null hypothesis the test statistic to be constructed is:



where $\mu = \mu_0$, σ is the population standard deviation, n is the sample size and $\bar{x}$ is the sample mean.

In many real life situations, the population standard deviation remains unknown. To overcome such type of situations, in case of large sample ($n \geq 30$), the unknown population standard deviation σ can be replaced by a sample standard deviation (s).

For the comparison of the calculated value of Z with the critical value at the given level of significance, Table 13.1 is used and on the basis of the comparison decision to accept or reject the null hypothesis is taken as per the procedure discussed in 13.6.

We will now discuss the following examples.

Example 1: A business firm conducted a sample survey 15 years ago and claimed that the average household income of a particular geographic region is Rs. 10,000. Mr. Baruah, recently appointed as

vice president in the firm has expressed doubts about the accuracy of the data. In order to verify the claim, a random sample of 200 households is tested. It is found that the mean household income is Rs. 11,000. Assume that the population standard deviation of the household income is Rs. 1200. Can we support the claim at 0.05 level of significance?

Solution: Here first we set up the null hypothesis H_0 and alternative hypothesis H_1 as follows:

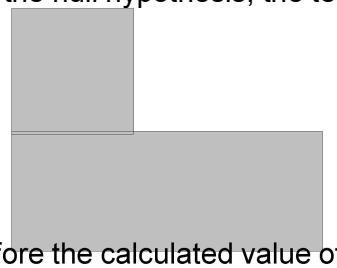
$H_0 : \mu = 10{,}000$

$H_1 : \mu \neq 10{,}000$

Given, $n = 200$, $\bar{x} = 11{,}000$, $\sigma = 1200$ and $\alpha = 0.05$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by



Therefore the calculated value of Z is 11.79

Again the critical value of Z at 5% level of significance for two-tailed test =1.96 Since the calculated value is Z greater than the critical value of Z, therefore the statistical decision is to reject the null hypothesis and conclude that the claim is not valid. Mr. Baruah's doubt about average household income was right.

Example 2: A company conducted a survey in the past to its consumers which reveals that the average household income is Rs. 20,000 in a particular region. Due to the rapid growth of development the company thinks that the average income may have changed. To verify this, a random sample of 50 households is selected and found that sample mean is Rs. 30,000 with a standard deviation of Rs. 15,000. Use 5% level of significance to verify the finding.

Solution: In the light of the problem, the null hypothesis H_0 and

alternative hypothesis H_1 are given as

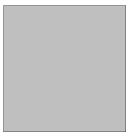
$$H_0 : \mu = 20{,}000$$

$$H_1 : \mu \neq 20{,}000$$

Given, $n = 50$, $\bar{x} = 30{,}000$, $s = 15{,}000$ and $\alpha = 0.05$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by





(since the population S.D. σ is unknown, therefore it can be replaced by the sample S.D. s)



Therefore the calculated value of Z is 4.714

Again the critical value of Z at 5% level of significance for two-tailed test =1.96 Since the calculated value Z is greater than the critical value of Z therefore the statistical decision is to reject the null hypothesis and conclude that average household income is significantly different from Rs. 20,000.

Example 3: The mean life of a sample of 80 fluorescent electric bulbs produced by a company is found to be 2000 hours with a standard deviation of 400 hours. Test the hypothesis that the mean life time of the bulbs produced in general is higher than the mean

life of 1900 hours at $\alpha = 0.01$ level of significance.

Solution: For the problem, let us take the null hypothesis that mean life of bulbs is not more than 1900 hours. Then the null hypothesis H_0 and alternative hypothesis H_1 will be as follows

$$H_0 : \mu \leq 1900$$

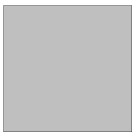
$$H_1 : \mu > 1900 \text{ (Right-tailed test)}$$

Given, $n = 80$, $\bar{x} = 2000$, $s = 400$ and $\alpha = 0.01$

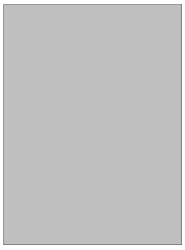
Since the sample size is large (≥ 30), therefore the test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by





(since the population S.D. σ is unknown, therefore it can be replaced by the sample S.D. s)



Therefore the calculated value of Z is 2.23

Again the critical value of Z at 1% level of significance for right-tailed test =2.33

Since the calculated value Z is less than the critical value of Z, therefore the statistical decision is to accept the null hypothesis. Hence we conclude that mean life of bulbs produced by the company is not more than 1900 hours.

13.7.2 Hypothesis Testing for Single Population Proportion

There are many areas where information that is highly dynamic in nature is generally expressed in terms of proportions or percentage such as, we often find that the market share of a company is 40% or 25% of the customers have switched from one brand to another brand or 10% of the items are defective. Quality defects, consumer preferences, market share etc. are some of the common areas where researchers may be interested to check the hypothesis whether such proportions have changed. For this a random sample of size n is selected from a large population possessing a particular attribute of interest (also termed as success) X then

$p =$ observed sample proportion of successes = 

Here we are interested to test the hypothesis that the proportion P in the population has a specified value P_0.

We set up the null hypothesis $H_0 : P = P_0$

Depending upon the nature, the alternative hypothesis may be any one of the following:

(i) $P \neq P_0$ (Two-tailed test)
(ii) $P > P_0$ (Right-tailed test)
(iii) $P < P_0$ (Left-tailed test)

Since n is large, therefore under the null hypothesis, the test statistic to be constructed is



where $P = P_0, Q = 1 - P = 1 - P_0$

To compare the calculated value of Z with the critical value at the given level of significance for the statistical decision, we use the Table 13.1.

Example 4: The production manager of a company that manufactures electric heaters claims that at least 10% of the heaters are defective. In order to test his believe, he takes a random sample of 100 heaters and finds that 12 heaters are defective. Use 5% level of significance to test the hypothesis.

Solution: Given $n = 100$, Number of defective heaters $= 12$

$p =$ sample proportion of defective heaters 

Null hypothesis $H_0 : P = 0.10$

Alternative hypothesis $H_1 : P \neq 0.10$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by





$\therefore$ The calculated value of $Z = 0.67$

The critical value of Z for two-tailed test at 5% level of significance=1.96 Since the calculated value is less than the critical value (0.67<1.96) therefore the statistical decision leads to acceptance of null hypothesis and conclude that the production manager's claim that at least 10% of the products are defective is valid. Consequently, the manufacturer can plan a strategy considering the fact that 10% of the products may be defective.

Example 5: A manufacturer of LCD TV claims that it is becoming quite popular in Guwahati, and that about 5% homes are having LCD TV. However, a dealer of conventional TVs claims that the percentage of homes with LCD TV is less than 5%. A sample of 500 household is surveyed, and it is found that only 18 households have LCD TV. Test at 1% level of significance whether the claim of the company is tenable.

Solution: Given, Given $n = 400$, Number of households having LCD TV

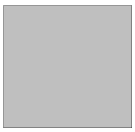
$p =$ sample proportion of LCD TV owners 

Null hypothesis $H_0 : P = 0.05$

Alternative hypothesis $H_1 : P < 0.05$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by:





$\therefore \quad |Z| = 0.515$

Therefore the calculated value of $|Z| = 0.515$

The critical value of Z for one-tailed test at 1% level of significance = 2.33 Since the calculated value is less than the critical value (0.515 < 2.33) therefore the statistical decision leads to the acceptance of null hypothesis and conclude that the manufacturer's claim that about 5% homes in Guwahati are having LCD TV is valid.

Example 6: The manager of a workshop claimed that at least 95%

of the equipments which he supplied to a factory conformed to specifications. To verify his claim a sample of 200 pieces of equipments was selected and revealed that 18 were faulty. Test his claim at a significance level of 0.05.

Solution: Given n=200,

$P =$ Proportion of pieces conforming to specification in the lot = 95% =0.95

$Q =$ Proportion of pieces not conforming to specification in the lot = $1 - P = 0.05$

Sample proportion 

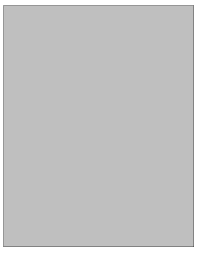
We take the null hypothesis that the proportion of pieces conforming to specification is 95% i.e., $H_0 : P = 0.95$

Alternative hypothesis $H_1 : P < 0.10$ (left-tailed test)

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by





$\therefore |Z| = 2.6$

The critical value of $|Z|$ at $\alpha = 0.05$ for one-tailed test = 1.645.

Since the calculated value of $|Z|$ is greater than the critical value

(i.e., 2.6 >1.645) therefore we reject the null hypothesis at 5% level and conclude that the claim of the manager is not valid.

[Alternatively, without taking the modulus we may compare the calculated value of Z with the critical value of $Z = -1.645$ at $\alpha = 0.05$ for left-tailed test. Since $-2.6 < -1.645$, the null hypothesis is to be rejected and thus both types of comparison lead us to the same decision.]



CHECK YOUR PROGRESS

Q 4: State whether the following statements are true or false:

(i) A tentative assumption about a population parameter is called the null hypothesis.

(ii) The equality part of the expression (either $\geq, \leq,$ or $=$) always appears in the null hypothesis.

(iii) The estimated variance of sampling distribution of a statistic is called standard error.

(iv) The test statistic to test $\mu = \mu_0$ for large sample is z-test.

(v) The usual notation for the standard error of the sampling distribution is σ/n.

(vi) Suppose for a particular hypothesis $\alpha = 0.05$ and $\beta = 0.10$. Then the power of this hypothesis test is 0.90.

(vii) If 2.81 is the critical value of test-statistic z, then the significance level of the test is 0.05.

(viii) The large sample test for testing $P = P_0$ for normal population is t-test.

(viii) For test of hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, the critical region at $\alpha = 0.05$ and $n \geq 30$, is $z > 1.96$.

(ix) When using the sample proportion $\bar{p}$ to test the hypothesis, the standard error of $\bar{p}$ is .

(x) Sampling distribution will be approximately normal if sample size is small.

P – Value Approach for Hypothesis Testing: This approach is applied in hypothesis testing for large samples and sometimes referred to as the observed level of significance. This approach is

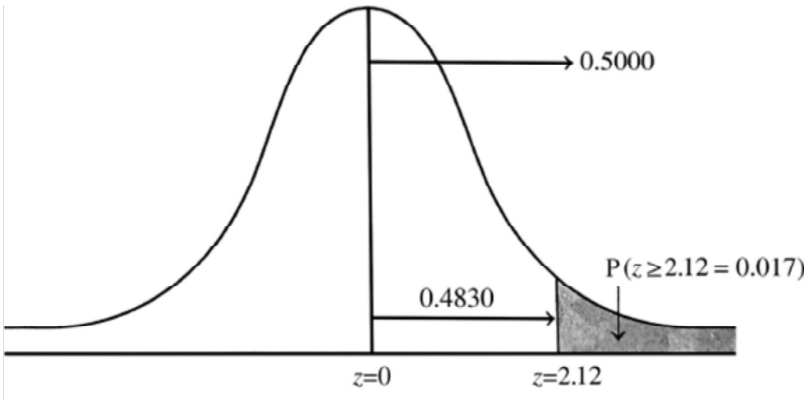
very advantageous especially after the introduction of many statistical software packages. The $P-$ value defines the smallest value of α for which the null hypothesis can be rejected. The decision rule for accepting or rejecting a null hypothesis based on the $P-$ value as discussed below:

Reject the null hypothesis H_0 when the value is $<\alpha$, otherwise, accept the null hypothesis H_0

For example, suppose a researcher conducting a hypothesis test with rejection region on the right tail of the distribution obtains an observed test statistic value of 2.12. From the normal table, the corresponding probability area for z value 2.12 is 0.4830. Therefore, the probability of getting a value greater than or equal to 2.12 is 5.0000-0.4830=0.017 (as shown in the following figure)

In this case, the p value is 0.017. For α =0.05 and α=0.01, this value falls under the rejection region (α=0.05, 0.017< 0.05 and α=0.01, 0.017< 0.01), so the null hypothesis will be rejected at 0.05 0.01 levels of significance. At α=0.01, the researcher cannot reject the null hypothesis for the value of equal to 0.017 because α=0.01< 0.017.

Figure 13.6: Probability of p (z is greater than or equal to 2.12)



The procedure of calculating the p value for a two-tailed test is slightly different. In a two-tailed test, the rejection region falls in both the tails of the normal distribution. For example, at α=0.05, the rejection region is located in both the tail areas of the distribution

in terms of 0.025% (0.025 +0.025=0.05) area in both the tails of the distribution. The researcher compares this value to the computed value for accepting or rejecting the null hypothesis. For this comparison, instead of splitting the value, we double the value and then compare that value with the value. In the above mentioned example, the researcher conducted a hypothesis test with rejection region on both the tails of the distribution and obtained the observed test statistic as 2.12 and corresponding value as 0.017. So, instead of splitting value in two parts, the researcher should double the value, i.e., (0.017×2=0.034). So, for a two-tailed test, this value (0.034) is compared with the values 0.05 and 0.01. Hence, at α =0.05 and α =0.01 the null hypothesis will be rejected but at α =0.01 the null hypothesis will be accepted (0.01< 0.034).

Example 7: A marketing research firm conducted a survey 10 years ago and observed that the average household income of a particular geographic region is Rs. 10,000. Mr. Gupta, who has recently joined the firm as a vice president has expressed doubts about the accuracy of the data. For verifying the data, the firm has decided to take a random sample of 200 households that yield a sample mean (for household income) of Rs. 10,200. Assume that the population standard of the household income is Rs. 1200. Use the p -value approach to verify Mr. Gupta's doubts. (Use $\alpha = 0.01$ as the level of significance.)

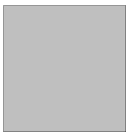
Solution: We set up the null and alternative hypothesis given by

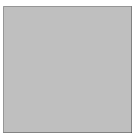
Null hypothesis $H_0 : \mu = 10{,}000$

Alternative hypothesis $H_1 : \mu \neq 10{,}000$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by





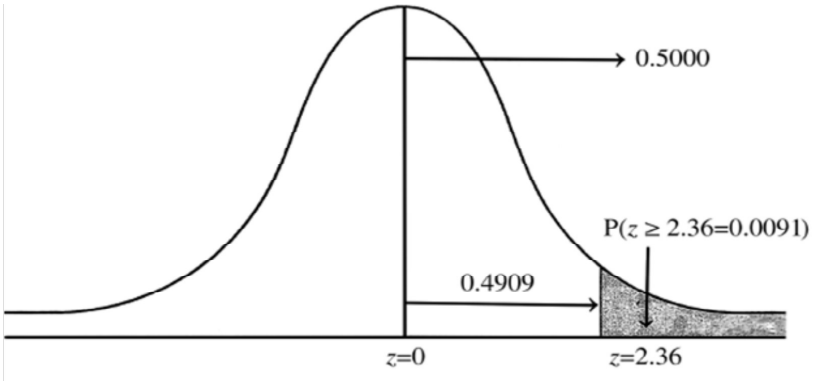
(since the population S.D. σ is unknown, therefore it can be replaced by the sample S.D. s)



Therefore the observed test statistic is computed as 2.36.

From the normal table, the corresponding probability area for Z value = 2.36 is 0.4909. So, the probability of obtaining a Z value greater than or equal to 2.36 is 0.5000-0.4909=0.0091 (as obtained from the following figure).

Figure 13.7: Probability of p (z is greater than or equal to 2.36)



For a two-tailed test, the obtained area value 0.0091 is to be multiplied by 2 and thus producing the value 0.0091×2=0.0182. Now since, $0.01 < 0.0182$, therefore we may accept the null hypothesis.

13.7.3 Hypothesis Testing for Difference between Two Population Means

In the real-world business environment, business researchers often encounter situations where they need to compare two means taken from two different samples drawn from two populations. On the

basis of the samples researchers make an inference about the difference in the means of the two populations. For example, suppose a researcher is interested to conduct a survey to study the consumer satisfaction for a particular brand of product in two different cities, Guwahati and Dibrugarh. In order to accomplish this, the researcher selects two different random samples from the two cities taken in the study. By using the procedures of testing of hypothesis, finally, the researcher draws a statistical conclusion regarding the consumer satisfaction in the two cities. Now, let $\overline{x_1}$ and $\overline{x_2}$ be the means of two large random samples of sizes n_1 and n_2 drawn from two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. We want to test the hypothesis whether the means of the two populations are equal or if there exists any significant difference between the population means.

Thus the null hypothesis is $H_0 : \mu_1 = \mu_2$

The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$

Assuming the null hypothesis to be true, the test statistic



The decision rule is same as discussed in the earlier units.

Example 8: An investigation of the relative merits of two kinds of flashlight batteries showed that a random sample of 100 batteries of brand A lasted on the average for 35.5 hours with a standard deviation of 1.8 hours, while a random sample of 80 batteries of brand B lasted on the average 36.8 hours with a standard deviation of 1.5 hours. Do the two brands differ significantly in quality? Use $\alpha = 0.05.$

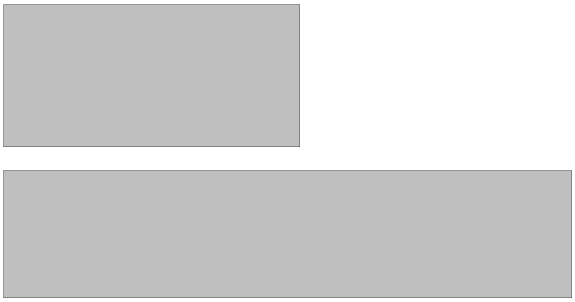
Solution: Let μ_1 and μ_2 denote the means of two populations.

Let us set up the null hypothesis that the two brands do not differ significantly in quality, i.e., $H_0 : \mu_1 = \mu_2$

and an alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by



$\therefore |Z| = 1.2195$

The critical value of $|Z|$ for two-tailed test at $\alpha = 0.05$ is 1.96.

Since the calculated value of $|Z|$ is less than the critical value of $|Z|$ therefore we may accept the null hypothesis and conclude that there is no significant difference between the average life times of the two brands of batteries.

Example 9: The means of two large samples of size 1000 and 2000 are found to be 67.5 and 68.0 respectively. Test the equality of the two populations each with S.D. 2.5.

Solution: Let us set up null and alternative hypothesis as:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Where μ_1 is the mean of the first population and μ_2 is the mean of the second population.

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by





$$\therefore |Z| = 5.16$$

Since the calculated value of $|Z|$ is greater than the critical value of $|Z|$ therefore we reject the null hypothesis and conclude that there is significant difference between population means.



CHECK YOUR PROGRESS

Q 5: A sample of 40 observations is selected from one population with S.D. (σ) 5. The sample mean is 102. A sample of 50 observations is selected from a second population with S.D. (σ) 6. The sample mean is 99. Conduct the following test of hypothesis (use $\alpha = 0.05$)

(a) Is this a one tailed or a two-tailed test? (b) State the decision rule. (c) Compute the value of the test statistic. (d) What is your decision regarding H_0?

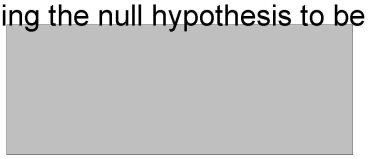
13.7.4 Hypothesis Testing for Difference between Two Population Proportions

At times, we come across certain situations where we are interested to study whether or not there exist significant differences between the proportions of two groups of, say, consumer in respect of a certain activity. Let p_1 and p_2 be the proportions of a certain attribute in two large random samples of sizes n_1 and n_2 drawn from two populations having proportions P_1 and P_2. On the basis of the samples, we want to test whether the two population proportions are equal or if there exist any significant difference between the

population proportions. Thus we will have null and alternative hypotheses as:

$$H_0 : P_1 = P_2$$
$$H_1 : P_1 \neq P_2$$

Assuming the null hypothesis to be true, the test statistic



where $Q_1 = 1 - P_1, \quad Q_2 = 1 - P_2.$

The procedure of taking decision is same as discussed earlier.

Note: (i) If P_1 and P_2 remains unknown, then they can be estimated by the corresponding sample proportions p_1 and p_2.

(ii) Sometimes, population proportion can be estimated by pooling the two sample proportions p_1 and p_2 into a single proportion p as given



In this case, the test statistic Z can be restated as



Example 10: In two populated states there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Solution: Given $n_1 = 1200, \quad n_2 = 900$

$$P_1 = 30\% = 0.30 \Rightarrow Q_1 = 0.70; \quad P_2 = 25\% = 0.25 \Rightarrow Q_2 = 0.75$$

Now we set up the null hypothesis H_0 that the difference in population proportions is likely to be hidden in samples i.e.,

$$H_0 : P_1 = P_2$$

And the alternative hypothesis is

$H_1 : P_1 \neq P_2$

Since the samples are large, under H_0 the test statistic Z is given by







Therefore the calculated value of Z =2.55

The critical value of Z at 5% level for two tailed test=1.96

Since the calculated value of Z > the critical value of Z, We reject the null hypothesis and hence, conclude that the difference in population proportions is not likely to be hidden in these samples.

Example 11: A Company has the head office at Calcutta and a branch at Guwahati. The personnel director wanted to assess if the workers at the two places would like the introduction of a new plan of work to enhance the quality of services and a survey was conducted for the purpose. Out of a sample of 500 workers at Calcutta 62% favoured the new plan. At Guwahati out of a sample of 400 workers 41% were against the new plan. Is there a significant difference between the two groups in their in their attitude towards the new $n_2 = 400, \quad p_2 = 59\% = 0.59 \Rightarrow q_2 = 1 - 0.59 = 0.41$ plan at 5% level?

Solution: Given

$n_1 = 500, \quad p_1 = 62\% = 0.62 \Rightarrow q_1 = 1 - 0.62 = 0.38$

Now we set up the null hypothesis that the population proportion of workers favouring the new plan in Calcutta and Guwahati is same, i.e. $H_0 : P_1 = P_2$,

The alternative hypothesis is

Since the samples are large, under the test statistic is given by



Where 





Therefore the calculated value of Z =0.917

The critical value of Z at 5% level for two tailed test=1.96

Since the calculated value of $Z <$ the critical value of , We may accept the null hypothesis and hence, conclude that there is no significant difference between the two groups in their attitude towards the new plan.



13.9 LET US SUM UP

- Hypothesis testing is a well-defined procedure that helps us to decide objectively whether to accept or reject the hypothesis based on the information available from the sample. Testing of hypothesis involves five different steps.
- In case of testing of statistical hypothesis, there are two types of hypothesis tests namely two-tailed test and one-tailed test. Two-tailed tests contain the rejection region on both the tails of the sampling distribution of a test statistic. As different from a two-tailed test, a one-tailed test contains the rejection region on one tail of the sampling

distribution of a test statistic.

- The testing of hypothesis involves two types of errors: Type I error and Type II error. A Type I error is committed by rejecting a null hypothesis when it is true. The possibility of committing Type I error is termed as level of significance and is denoted by α. On the other hand, a type II error is committed by accepting a null hypothesis when it is false. The likelihood of committing Type II error is denoted by β.
- Hypothesis testing can be performed by applying the approaches: the $z-$value approach and $p-$value approach. For testing hypothesis about a single population mean, z formula can be used if the sample size is large $(n \geq 30)$ for any population with known population standard deviation σ. But in practical situations if σ is not available then it can be estimated by a sample standard deviation s. The p value defines the smallest value of for which the null hypothesis can be rejected.
- Hypothesis can also be performed for single population proportion as well as for the difference in two population proportions. In case of two random samples, on the basis of the difference in sample proportions, a researcher can estimate the difference in population proportions.

13.10 FURTHER READING

- Gupta S.C. (1994). Fundamentals of statistics. New Delhi, Himalayan Publishing House.
- Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.
- Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.
- Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.

13.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: Hypothesis testing is the process of testing a statement or belief about a population parameter by the use of information collected from a sample (s).

Ans to Q No 2: Type I error is the probability of rejecting the null hypothesis when it is true. It is denoted by Greek letter α .

Ans to Q No 3: Type II error is the probability of accepting the null hypothesis when it is false. It is denoted by Greek letter β .

Ans to Q No 4: (i) T, (ii) T, (iii) F, (iv) T, (v) F, (vi) T, (vii) F, (viii) F, (ix) T, (x) F.

Ans to Q No 5: (a) Two-tailed test, (b) Reject H_0 $|z| > 1.96$,

(c) H_0

13.12 MODEL QUESTIONS

Q 1: Discuss the various steps involved in testing of hypothesis.

Q 2: What do you understand by null hypothesis, alternative hypothesis and level of significance?

Q 3: Define Type I error and Type II error.

Q 4: What is meant by one-tailed test and two-tailed test?

Q 5: The mean breaking strength of the cables supplied by a manufacturer is 1800 with a standard deviation of 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance?

Q 6: The MacBurger restaurant chain claims that the mean waiting time of customers is 3 minutes with a population standard deviation of 1 minute. The quality- assurance department found in a sample of 50

customers at the Warren Road MacBurger that the mean waiting time was 2.75 minutes. At the 5% significance level, can we conclude that the mean waiting time is less than 3 minutes?

Q 7: A company manufacturing a certain type breakfast cereal claims that 60% of all housewives prefer that type to any other. A random sample of 300 housewives contains 165 who do prefer that type. At 5% level of significance, test the claim of the company.

Q 8: A man buys 50 electric bulbs of 'Philips' and 50 electric bulbs of 'HMT'. He finds that 'Philips' gave an average life of 1500 hours with a standard deviation of 60 hours and 'HMT' bulbs gave an average life of 1512 hours with a standard deviation of 80 hours. Is there a significant difference in the mean life of the two makes of bulbes?

Q 9: 500 units from a factory are inspected and 12 are found to be defective, 800 units from another factory are inspected and 12 are found to be defective. Can it be concluded that at 5% level of significance production at the second factory is better than in first factory?

*** ***** ***

UNIT 14: TEST OF SIGNIFICANCE

UNIT STRUCTURE

14.1 Learning Objectives

14.2 Introduction

14.3 The Chi-Square (χ^2) Distribution

14.4 Applications of χ^2 Test

- 14.4.1 χ^2 Test of Goodness of Fit
- 14.4.2 χ^2 Test for Independence of Attributes
- 14.4.3 χ^2 Test for a Specified Population Variance
- 14.4.4 χ^2 Test of Homogeneity

14.5 Testing of Hypothesis in case of Small Samples (t-Test)

- 14.5.1 Characteristics of t-distribution
- 14.5.2 Applications of t-distribution
- 14.5.3 Hypothesis Testing for Single Population Mean
- 14.5.4 Hypothesis Testing for Difference between Two Population Means
- 14.5.5 Paired t-test for Difference between Two Population Means

14.6 Let Us Sum Up

14.7 Further Reading

14.8 Answer to Check Your Progress

14.9 Model Questions

14.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- introduce the Chi-square distribution for the comparison of more than two means or proportions
- use the Chi-square distribution to test whether two classifications of the same data are independent of each other
- use the Chi-square distribution for testing the hypothesis about a single population variance
- use t-test in small samples.

14.2 INTRODUCTION

In the previous unit we have learnt some statistical tests (using data

from either one or two samples) that are used to test the hypothesis about the population parameters, with a number of business related examples. These tests are based on the assumption that the samples are drawn from a normally distributed population or, more appropriately, that the sample means are normally distributed. Such tests are known as parametric tests as the testing procedure involves the assumption about the type of population of parameters. On the other hand, in some cases, the researcher may not be sure of whether the population distribution is normal. The statistical tests that do not require prior knowledge about the population are termed as non-parametric tests. Chi-square test is one of the important non-parametric methods.

A non-parametric method, also known as distribution free test satisfies at least one of the following criteria:

(i) The method does not take into consideration any population parameter.

(ii) It is applied only on categorical data that are non-numerical and frequency counts of categories for one or more variables.

(iii) It does not depend on the form of the underlying population distribution, in particular, the requirement of normality.

In this unit we will focus on χ^2 (Chi-square) test. This test enables us to test whether more than two population proportions are equal. In fact, it allows to do a lot more than just test for the equality of several proportions. If we classify a population into several categories with respect to two attributes (such as age and job performance), we can then employ a chi-square test to test whether the two attributes are independent of each other.

14.3 THE CHI-SQUARE (χ^2) DISTRIBUTION

The symbol χ stands for the Greek letter 'chi'. This test was first developed by Karl Pearson in testing of hypothesis in 1900. The sampling distribution of χ^2 is called χ^2 distribution. This distribution has only one parameter called the degrees of freedom (df) and the shape of the curve depends on the degrees of freedom. The distribution is skewed to the right. Being a square quantity, χ^2 distributi3on can never be negative. In fact, χ^2 is a continuous probability distribution with range 0 to ∞ (Figure 14.1)

The probability density function of distribution is given by

where ν is is the degrees of freedom, C is a constant depending upon the degrees of freedom, and $e = 2.71828.$

χ^2 test statistic can be defined as below:

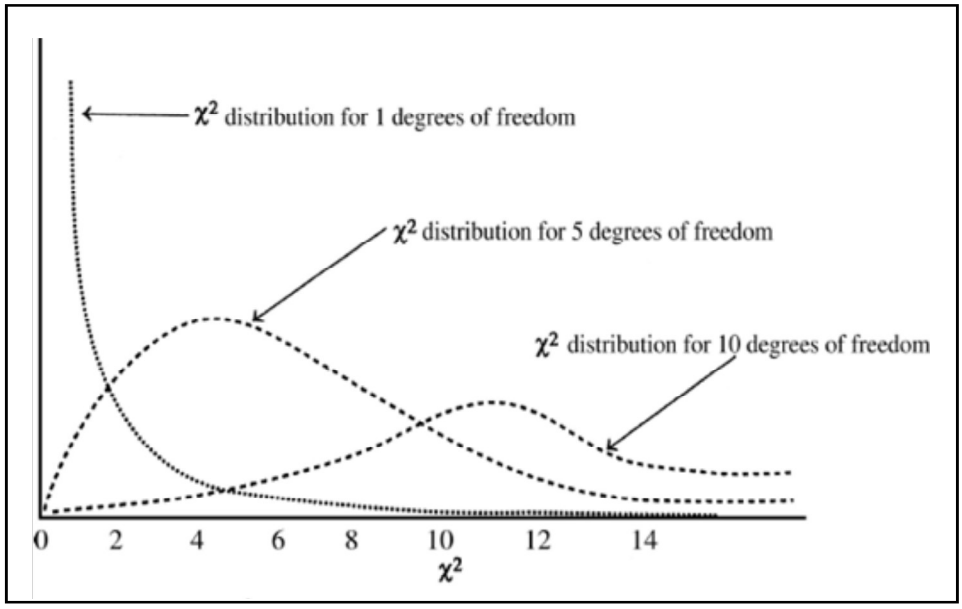
with $df = k - 1 - c$

where O is the observed frequency, E the expected frequency, k the number of categories, and c the number of parameters being estimated from the sample data.

The computed value of χ^2 is compared with the critical value of χ^2 at a specified level of significance and degrees of freedom. The decision to accept or reject the null hypothesis is given by

If $\chi^2_{cal} < \chi^2_{critical}$ then we may accept the null hypothesis, otherwise we reject the null hypothesis.

Figure 14.1: χ^2 Distribution with 1, 5 and 10 Degrees of Freedom



Properties of Chi-square test:

The following properties are useful while using χ^2 test statistic to analyze its sampling distribution:

(i) The density function of χ^2 does not involve any parameter and the shape of the curve depends upon the degrees of freedom only. The distribution becomes more and more symmetrical as the degrees of freedom increase and tends to normal distribution for high value of degrees of freedom ($\nu \geq 30$).

(ii) Mean = Degrees of freedom = ν, mode= $\chi^2 - 2$ and variance= 2ν

(iii) The values of χ^2 lie between 0 and ∞.

(iv) If $x_1, x_2, ..., x_n$ be a random sample drawn from a normal population with mean μ and variance σ^2, then

(a)  χ^2 -distribution with n degrees of freedom.

(b)  follows χ^2 -distribution with degrees of freedom $n-1$, $\bar{x}$ being the sample mean.

(v) For large degrees of freedom (ν), the standard variate



may approximately be considered as a standard normal variate.

Conditions for applications of Chi-square test:

The following conditions need to be satisfied before applying χ^2 as a test statistic for hypothesis testing:

(i) In a contingency table, an expected frequency of less than 5 in a cell is less than the frequency required to apply the test. In such cases we apply the χ^2 technique of 'pooling' which consists in adding the may approximately be considered as a standard normal variate.

Conditions for applications of Chi-square test:

The following conditions need to be satisfied before applying χ^2 as a test statistic for hypothesis testing:

(i) In a contingency table, an expected frequency of less than 5 in a cell is less than the frequency required to apply the test. In such cases we apply the technique of 'pooling' which consists in adding the frequencies which are less than 5 with the preceding or succeeding frequency, so that the resulting sum is 5 or more and adjust for the degrees of freedom accordingly.

(ii) The sample should consist of at least 50 observations and should be drawn randomly from the population.

(iii) The data should be expressed in original units, rather than in percentage or ratio form.

(iv) The constraints on the cell frequencies, if any, should be linear such as $\sum O = \sum E = N$, the total frequency.

14.4 APPLICATIONS OF χ^2 TEST

The Chi-square test is not based on assumption of normality and can be used to both large and small samples. Chi-square distribution is a widely used since it can be employed in situations in which the normal distribution cannot be applied. Chi-square distribution has a number of applications, some of which are enumerated as below:

(i) to test goodness of fit,

(ii) to test the independence of attributes,

(iii) to test if the population has a specified value of standard deviation or variance.

(iv) to test if two or more populations are homogenous with respect to some characteristics of interest.

Critical Values of χ^2 :

From the χ^2 table (given in the Appendix) it may be observed that that the tabulated (critical) values of χ^2 increase as the degrees of freedom increases and the level of significance decreases.

Let us define the symbol $\chi_n^2(\alpha)$ as the value of χ^2 for $n\ d.f.$ such that the area to the right of this point is α, i.e.,



is the rejection region. This is shown in the Figure 14.2.

Figure 14.2: Critical Value and Rejection region of χ^2 Statistic

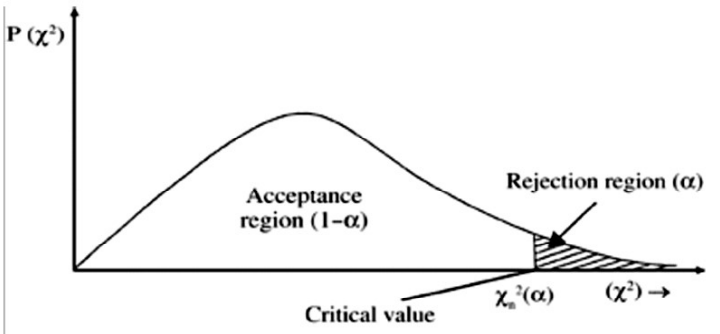


Fig. 10.2 : Critical Value and rejection region of χ^2 statistic



CHECK YOUR PROGRESS

Q 1. State whether the following statements are true or false:

(i) The critical value of a statistic is the value which cuts off the region for the rejection of null hypothesis.

(ii) The χ^2-test requires that the sampling distribution should be normally distributed.

(iii) If null hypothesis is accepted, then it is concluded that data analysis did not produce significant results.

(iv) In order to reject the null hypothesis $\chi^2_{cal} > \chi^2_{table}$

(v) The sampling distribution is a family of curves, the distribution of which varies with the degrees of freedom.

Now we discuss various applications of χ^2.

14.4.1 χ^2 Test of Goodness of Fit

χ^2 test enables us to ascertain whether the known probability distributions such as Binomial, Poisson and Normal distributions fit with an observed sample observation. Thus χ^2 test for goodness of fit, a statistical test, provides a platform that can be used to ascertain whether theoretical probability distributions coincide with empirical sample distributions. This test compares the theoretical (expected) frequencies with the observed (actual) frequencies to obtain the deviation between theoretical and observed frequencies. For the application of χ^2 test, first a theoretical distribution is hypothesized for a given population and then the test is carried out to make sure whether or not the sample data could come from the population of interest with the hypothesized theoretical distribution. The steps used for hypothesis testing can also be performed using the χ^2 goodness of fit test.

Steps to conduct χ^2 test for goodness of fit: The general steps to conduct a goodness of fit test for any hypothesized population distribution are summarized as below:

Step 1: Set up the null and alternative hypotheses as

Null hypothesis H_0 : No difference between the observed and expected sets of frequencies.

Alternative hypothesis H_1 : There is a difference.

Step 2: Calculate expected frequencies (E values) in each category by multiplying the category probability by the sample size ().

Step 3: Compute the test statistic



where O is the observed frequency, E the expected frequency.

Step 4: Using a specified level of significance α and $d.f. = n - 1$ provided that the number of expected frequencies are 5 or more for all categories, determine the critical value of χ^2 .

Step 5: Compare the computed value of χ^2 with the critical value of χ^2 at a specified level of significance and degrees of freedom. The decision to accept or reject the null hypothesis is given by

If χ^2_{cal} is less than the critical value $\chi^2_{\alpha,n-1}$ then we may accept the null hypothesis, otherwise we reject the null hypothesis.

Example 1: A company is concerned about the increasing violent altercations between its employees. The number of violent incidents recorded by the management during six randomly selected months is given by

Months	Number of violent incidents
Jan	55
Feb	65
Mar	68
Apr	72
May	80
Jun	85

Test whether the data fits a uniform distribution. (Use $\alpha = 0.05$)

Solution: Let us set up null and alternative hypotheses as

Null hypothesis H_0 : Number of violent altercations are uniformly distributed over the months.

Alternative hypothesis H_1 : Number of violent altercations are not uniformly distributed over the months.

In order to test the null hypothesis, the appropriate test statistic is given by



where O is the observed frequency, E the expected frequency.

Now the expected frequencies will be calculated by dividing total observed frequencies by number of

months. Hence, the expected frequency 

Now to compute the test statistic we prepare the following table:

Months	O	E	$\frac{(O-E)^2}{E}$
Jan	55	70	3.2142
Feb	65	70	0.3571
Mar	68	70	0.0571
Apr	72	70	0.0571
May	80	70	0.9142
Jun	85	70	2.0571
Total	420		6.65

Thus $\chi^2 = 6.65$

$\therefore$ The calculated value of χ^2 is 6.65

The degrees of freedom= $n-1=6-1=5$

The critical value of χ^2 at 5% level of significance for 5 degrees of freedom i.e., $\chi^2_{0.05,5} = 11.07$

Since the calculated value of is less than the critical value , therefore we may accept the null hypothesis and conclude that the number of violent altercations is uniformly distributed over the months.

Example 2: A brand manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey in which the country was divided into four geographical regions, a random sampling of 100 consumers in each region was surveyed, with the following information:

	Region				
	NE	NW	SE	SW	Total
Purchase the brand	40	55	45	50	190
Do not purchase	60	45	55	50	210
Total	100	100	100	100	400

Develop a table of observed and expected frequencies for this problem.

(i) Calculate the sample Chi-square value.

(ii) State the null and alternative hypothesis.

(iii) At $\alpha = 0.05$, test whether brand share is the same across the four regions.

Solution: To calculate the expected frequencies for the corresponding observed frequencies, we apply the following formula

(Row total × Column total)/Grand total.

The calculations of expected frequencies for the corresponding observed frequencies are given below:

		O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Row 1	NE	40	47.5	-7.5	56.25	1.18
	NW	55	47.5	7.5	56.25	1.18
	SE	45	47.5	-2.5	6.25	0.13
	SW	50	47.5	2.5	6.25	0.13
Row 2	NE	60	52.5	7.5	56.25	1.07
	NW	45	52.5	-7.5	56.25	1.07
	SE	55	52.5	2.5	6.25	0.12
	SW	50	52.5	-2.5	6.25	0.12
Total						5.00

(i) The sample Chi-square statistic is given by



where O is the observed frequency, E the expected frequency.

$\therefore$ From the above table we have,

$$\chi^2 = 5.00$$

(ii) We state the null and alternative hypotheses as:

Null hypothesis H_0 : The brand share is evenly distributed.

Alternative hypothesis H_1 : The brand share is not evenly distributed.

(iii) The calculated value of χ^2 is 5.00

The degrees of freedom = $(r-1)(c-1) = (2-1)(4-1) = 3$

The critical value of χ^2 at 5% level of significance for 3 degrees of freedom i.e., $\chi^2_{0.05,3} = 7.815$

Since the calculated value of is less than the critical value, therefore we may accept the null hypothesis and conclude that the brand share is evenly distributed in all the four regions of the country.

Example 3: A survey of 800 families with 4 children each revealed the following distribution:

Test whether the data are consistent with the hypothesis that the binomial law holds and the chance of a male birth is equal to that of a female birth.

Solution: Here we set up the null hypothesis that the data are consistent with the binomial law of equal probability for male and female births, i.e.,

, (where p denotes the probability of a male birth.)

Using the binomial probability law, the expected frequency $f(x)$ of x male births is given by:

Number of births		Number of families
Male	Female	
0	4	32
1	3	178
2	2	290
3	1	236
4	0	64

$f(x) = N.p(x) = N \times {}^{n}C_{x}\, p^{x}q^{n-x}$, where

$N = 800, n = 4, \; x = 0, 1, 2, 3, 4.$

$= 50 \times {}^{4}C_{x}$

……(*)

Thus putting $x = 0, 1, 2, 3, 4$ successively in (*), we can determine the expected binomial frequencies as given below:

Number of Male births (x)	Expected frequency $f(x)$
0	$50\times{}^{4}C_{0}=50\times1=50$
1	$50\times{}^{4}C_{1}=50\times4=200$
2	$50\times{}^{4}C_{2}=50\times6=300$
3	$50\times{}^{4}C_{3}=50\times4=200$
4	$50\times{}^{4}C_{4}=50\times1=50$
Total	800

Let us now apply χ^2 test to examine the goodness of fit of the given data to the above binomial distribution. Calculations are given below:

	Observed frequency(O)	Expected frequency (E)	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	32	50	-18	324	6.48
1	178	200	-22	484	2.42
2	290	300	-10	100	0.33
3	236	200	36	1296	6.48
4	64	50	14	196	3.92
Total	800	800	0		19.63

Thus $\chi^2 = 19.63$

$\therefore$ The calculated value of χ^2 is 19.63

The degrees of freedom= $5-1=4$

The critical value of χ^2 at 5% level of significance for 4 degrees of freedom i.e., $\chi^2_{0.05,4} = 9.488$

Since the calculated value of χ^2 is greater than the critical value $\chi^2_{0.05,4}$, therefore we reject the null hypothesis and conclude that the hypothesis of equal male and female births is not valid.

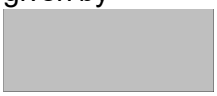
Example 4: The figures given below are (a) the observed frequencies of a distribution, and (b) the frequencies of the normal distribution having the same mean, standard deviation and total frequency as in (a). Apply the χ^2 test of goodness of fit.

(a)	1	12	66	220	495	792	924	792	495	220	66	12	1
(b)	2	15	66	210	484	799	944	799	484	210	66	15	2

Solution: Since we are given the observed frequencies and expected frequencies of the normal distribution, therefore we apply the χ^2 test of goodness of fit.

Now we set up the null hypothesis that normal distribution is a good fit to the given frequency distribution, i.e., the given observed frequencies and the expected normal frequencies do not differ significantly.

In order to test the null hypothesis, the appropriate test statistic is given by



where O is the observed frequency, E the expected frequency.

Now to compute the χ^2 test we prepare the following table:

	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
	1, 12	2, 15	-4	16	0.94
	66	66	0	0	0.00
	220	210	10	100	0.48
	495	484	11	121	0.25
	792	799	-7	49	0.06
	924	944	-20	400	0.42
	792	799	-7	49	0.06
	495	484	11	121	0.25
	220	210	10	100	0.48
	66	66	0	0	0.00
	12, 1	15, 2	-4	16	0.94
Total	4096	4096	0		3.88

Thus $\chi^2 = 3.88$

$\therefore$ The calculated value of χ^2 is 3.88

The degrees of freedom=13-5=8

(The 13 original classes have been reduced to 11 by grouping, thus 2 degrees of freedom are lost in pooling. In addition, the mean, the standard deviation and total frequency of the original distribution have been used in calculating the theoretical frequencies, thus introducing three restraints. The number of degrees of freedom is accordingly 8.)

The critical value of χ^2 at 5% level of significance for 8 degrees of freedom i.e., $\chi^2_{0.05,8} = 15.507$

Since the calculated value of χ^2 is much less than the critical value $\chi^2_{0.05,8}$, it is highly non-significant. Hence, we may accept the null hypothesis and conclude that the normal distribution is a good fit to the given distribution.



CHECK YOUR PROGRESS

Q 2: In a particular chi-square goodness-of-fit test there are four categories and 200 observations. Use the 5% significance level.

(a) How many degrees of freedom are there?

(b) What is the critical value of chi-square?

Q 3: The null hypothesis and alternative hypothesis are H_0 : The frequencies are equal and H_1 : The frequencies are not equal

Category	f
A	10
B	20
C	30

(a) State the decision rule ($\alpha = 0.05$), (b) Compute the value of Chi-square.

(c) What is your decision regarding?

14.4.2 χ^2 Test for Independence of Attributes

In many business situations, a market researcher might be interested to study the relationship between two variables or to check whether they are independent of each other. For example, an edible oil company may be interested in knowing whether the purchase of oil is independent of the customer's age or whether it is dependent on the customer's age. These are two different situations and consequently the company needs to frame a production and selling strategy. Similarly following are the situations where the Chi-square test of independence can be employed to analyse attributes with multiple categories.

- Whether university students classified by gender are independent of courses of study.
- Whether wage-earners classified by education level are independent of income.
- Whether a product produced is acceptable or not is independent of the shifts in which it was produced.

Thus when observations are classified on the basis of two qualitative variables or attributes and arranged in a table, the resulting table is known as a contingency table (Table 14.1). χ^2 test of independence uses this contingency table for determining independence of two qualitative variables.

Table 14.1: Contingency table

Variable B	Variable A				Total
	A_1	A_2	.....	A_k	
B_1	O_{11}	O_{12}	.....	O_{1k}	R_1
B_2	O_{21}	O_{22}		O_{2k}	R_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
B_r	O_{r1}	O_{r2}	.....	O_{rk}	R_r
Total	C_1	C_2		C_k	N

It is to be noted that the variables A and B have been classified into mutually exclusive categories. The value O_{ij} is the observed frequency for the cell in i[th] row and j[th] column. R_i is the total of the i^{th} row and C_j is the total of the j^{th} column. The row and column totals are added up to get a grand total (N), which represents the sample size. It is to be noted that to apply the Chi-square test we need to calculate the expected frequency.

The calculation of expected frequency, E_{ij} , corresponding to an observed frequency O_{ij} in row i and column j is based on the concept of multiplicative law of probability. Using this concept of probability the expected frequency, , is calculated as

$E_{ij} =$

Thus expected frequency for any cell can be computed by using the formula

The analysis of a two-way contingency table helps to determine whether the two variables are unrelated or independent of each other. Thus the null hypothesis for a Chi-square test of independence is that the two variables are independent.

Under the null hypothesis, the χ^2 test statistic is defined as

The degrees of freedom for the Chi-square test of independence are given by

Degrees of freedom= (Number of rows-1) × (Number of columns-1)

$$=(r-1)\times(c-1)$$

Example 5: A certain drug is claimed to be effective in curing colds. In an experiment on 500 persons with cold, half of them were given the drug and half of them were given sugar pills. The patients' reactions to the treatment are recorded in the following table:

	Consequence			
Treatment	Helped	Reaction	No effect	Total
Drug	150	30	70	250
Sugar pills	130	40	80	250
Total	280	70	150	500

On the basis of the data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills?

	Consequence			
Treatment	Helped	Reaction	No effect	Total
Drug	$\frac{250\times280}{500}=140$	$\frac{250\times70}{500}=35$	250-(140+O5)=75	250
Sugar pills	140	35	75	250
Total	280	70	150	500

Now we arrange the observed and expected frequencies in the following table in order to calculate the χ^2 test statistic.

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
150	140	10	100	0.714
130	140	-10	100	0.714
30	35	-5	25	0.714
40	35	5	25	0.714
70	75	-5	25	0.333
80	75	5	25	0.333
Total				3.522

Thus the χ^2 test statistic is



Number of degrees of freedom= (r-1) (c-1) = (2-1) (3-1) =2

The critical value of the χ^2 at 5% level of significance for 2 degrees

of freedom = 5.99

Since the calculated value of the is less than the critical value, therefore we may accept the null hypothesis. Hence conclude that drug is no better than sugar pills in curing colds.

Example 6: Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	Number of students in each level				Total
	Below average	Average	Above Average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

Solution: We state he null hypothesis that the data obtained are independent of the sampling techniques adopted by the two researchers.

Under the null hypothesis of independence, the expected frequencies are calculated as:



Now we arrange the observed and expected frequencies in the following table in order to calculate the χ^2 test statistic.

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.500
40	42	-2	4	0.095
33	31	2	4	0.129
25	23	2	4	0.174
2	4	-2	4	1.000
Total	300	0		2.097

Thus the χ^2 test statistic is



Number of degrees of freedom= (r-1)(c-1)=(4-1)(2-1)=3

The critical value of the χ^2 at 5% level of significance for 3 degrees of freedom = 7.82

Since the calculated value of χ^2 the is less than the critical value, therefore we may accept the null hypothesis. Hence conclude that the sampling techniques adopted by the two investigators do not differ significantly.



CHECK YOUR PROGRESS

Q 4: A social scientist sampled 140 people and classified them according to income level and whether or not they played a state lottery in the last month. The sample information is reported below. Is it reasonable to conclude that playing the lottery is related to income level? Use the 5% significance level.

	Income			
	Low	Middle	High	Total
Played	46	28	21	95
Did not play	14	12	19	45
Total	60	40	40	140

(a) What is this table called?

(b) State the null and alternative hypotheses.

(c) What is the degrees of freedom?

(d) What is the decision rule?

(e) Determine the value of χ^2.

(f) Make a decision about H_0.

14.4.3 χ^2 Test for a Specified Population Variance

Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2. We want to test the hypothesis that the population variance has a specified value.

We set up the null hypothesis and the alternative hypothesis as

$$H_0 : \sigma^2 = \sigma_0^2 \text{ and } H_1 : \sigma^2 > \sigma_0^2$$

Then under the null hypothesis, the test statistic



follows χ^2 distribution with $(n-1)\ d.f.$, where



, denotes the sample variance.

By comparing the calculated value of χ^2 with the tabulated value of for at the specified level of significance, we may accept or reject the null hypothesis at that level of significance.

Remarks: 1. test given by



can be applied only if the population is normal.

2. If the sample size is large, say, greater than 30, then we can use Fisher's approximation, viz.,

$$\sqrt{2\chi^2} \sim N\left(\sqrt{2n-1}, 1\right)$$

And then apply the normal test by considering the statistic

$$Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$$

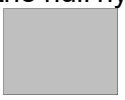
Example 7: A battery manufacturer claims that the lives of his batteries have a standard deviation of 0.9 year. Can we accept manufacture's claim if the lives of a random sample of 10 batteries have a standard deviation of 1.2 years? (Use a 5% level of significance)

Solution: We set up the null and alternative hypotheses as

Null hypothesis $H_0 : \sigma = 0.9\ years$

Alternative hypothesis $H_1 : \sigma > 0.9\ years$

Under the null hypothesis, the test statistic



~Chi-square distribution with $(n-1)\ d.f.$

Now 

Degrees of freedom= n-1 =10-1=9

The critical value of χ^2 at 5% level for 9 d.f. i.e., $\chi^2_{0.05}(9) = 16.92.$

Since the critical value of χ^2 is greater than the tabulated value of value of χ^2 at 5% level for 9 d.f., we reject the null hypothesis and conclude that the standard deviation of life of the batteries is more than 0.9 years.

Example 8: The weights in kg. of 10 students are given below:

38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that the variance of the distribution of weights of all students from which the above sample of 10 students was drawn, is equal to 20 square kg.?

Solution: We set up the null and alternative hypotheses as:

Null hypothesis $H_0 : \sigma^2 = 20\ sq.\ kg.$

Alternative hypothesis $H_1 : \sigma^2 > 20\ sq.\ kg.$

Under the null hypothesis, the test statistic



~Chi-square distribution with $(n-1)\ d.f.$

Now we prepare the following table to calculate the sample variance

Weight in kg. (x)	$(x-\bar{x})$ $x-47$	$(x-\bar{x})^2$
38	-9	81
40	-7	49
45	-2	4
53	6	36
47	0	0
43	-4	16
55	8	64
48	1	1
52	5	25
49	2	4
Total=470		280





Now, 

Degrees of freedom= n-1=10-1=9

The critical value of χ^2 i.e., $\chi^2_{0.05}(9) = 16.92$

Since the calculated value of χ^2 is less than the critical value of $\chi^2_{0.05}(9)$, therefore we may accept our null hypothesis and conclude that the variance of the distribution of weights of all students in the population is 20 sq. kg.

14.4.4 χ^2 Test of Homogeneity

χ^2 test of homogeneity is used to determine whether two or more independent variables are drawn from the same population or from different populations. In other words, χ^2 test of homogeneity is used to know whether two or more populations are homogeneous with respect to some characteristics of interest. For example, a researcher may be interested in knowing whether the employees from three departments, production, finance, and personnel, feel the same about the requirements of the top management in terms of hard work expected from employees. The amount of hard work can also be classified into three groups, namely, very hard, hard, and easy going. In this situation, we can state a null hypothesis that the opinion of all the groups is the same about the requirement of hard work. In other words, the null hypothesis assumes that the three classifications are homogeneous in terms of their opinion about the amount of hard work required by the top management.

It may be noted that, in both χ^2 test of homogeneity and χ^2 test of independence, one is concerned with the cross tabulation of the data and hypotheses testing procedure is also same. The differences that exist between the χ^2 test of homogeneity and χ^2

test of independence are: (i) In χ^2 test of independence, a researcher is interested in testing whether two attributes are independent. On the other hand in, χ^2 intest of homogeneity, a researcher investigates whether two or more populations are homogenous with respect to some characteristic of interest. (ii) in χ^2 test of homogeneity, two or more independent samples are drawn from each population whereas in case of χ^2 test of independence, a single sample is drawn from a population.

Example 9: A television company has launched a new product with some advanced features. The company wants to know the opinion of consumers about this product with respect to four characteristics: preferred brand with new features did not prefer brand with new features, preferred only a few new features, and indifferent. The company has divided consumers into three groups—executives/ officers; businessmen, and private consultants. It has taken a random sample of size 459 and obtained the following:

Consumers	Executives/Officers	Businessmen	Private consultants	Row total
Preferred brand with new features	35	25	40	100
Did not prefer brand with new features	30	45	34	109
Preferred only a few new features	45	50	25	120
Indifferent	25	55	50	130
Column total	135	175	149	459

Use χ^2 test of homogeneity and draw inference from the data.

Solution: Let us set up null and alternative hypotheses as

Null hypothesis H_0 : Opinion of all the groups is the same about the product with new features.

And alternative hypothesis H_1 : Opinion of all the groups is not the same about the product with new features.

In order to test the null hypothesis, the appropriate test statistic is given by:



where O is the observed frequency, E the expected frequency.

To calculate the expected frequencies for the corresponding observed frequencies, we apply the following formula

(Row total × Column total)/Grand total.

Now for the calculation of the expected frequencies against the corresponding observed frequencies and to compute the test statistic we prepare the following table:

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2/E$
35	29.41	5.59	1.0625
30	32.06	-2.06	0.1324
45	35.29	9.71	2.6717
25	38.24	-13.24	4.5841
25	38.13	-13.13	4.5213
45	41.56	3.44	0.2847
50	45.75	4.25	0.3948
55	49.56	5.44	0.5971
40	32.46	7.54	1.7514
34	35.38	-1.38	0.0538
25	38.95	-13.95	4.9962
50	42.20	7.8	1.4417
420			22.4917

Thus the χ^2 test statistic is



The number of degrees of freedom $= (r-1)(c-1) = (4-1)(3-1) = 6$

The critical value of the χ^2 at 5% level of significance for 6 degrees of freedom = 12.59

Since the calculated value of the χ^2 is greater than the critical value, therefore we reject the null hypothesis and conclude that the opinion of all the groups is not the same about the product with new features. Hence, the company has to consider different groups and their needs separately.

Precautions about using the Chi-square test: In order to apply Chi-square hypothesis test properly, one has to be extremely careful and keep in mind certain precautions. First, a sample should

be large enough. If the expected frequencies are too small, the value of Chi-square gets over estimated. This will result in the rejection of the null hypothesis in several situations. To overcome this problem, we must ensure that the expected frequency in any cell of the contingency table should not be less than 5. In case the expected frequency less than 5 in more than one cell, we can combine these to obtain the expected frequency of at least 5. Another point to note is that the calculations must be made with actual numbers in each cell and not with proportions or percentages. If the proportions or percentages were used, then the theoretical distribution would not be applicable.

When the calculated value of Chi-square turns out to be more than the critical value at a specified level of significance, we reject our null hypothesis. On the other hand, when Chi-square value is less than the critical value, we may accept our null hypothesis. However, when the Chi-square value turns out to be zero, we have to be extremely careful to confirm that there is no difference between the observed and the expected frequencies. Such a situation may sometimes arise on account of a faulty manner used in the data collection.

In most of the cases, the problems of Chi-square involve simple calculations. However, for large sets of data the Chi-square test involves very comprehensive calculations. In all such situations, computer should be used. Several computer Statistics packages contain routines for carrying out Chi-square tests.



CHECK YOUR PROGRESS

Q 5: When chi-square test of homogeneity is useful?

……

……

……

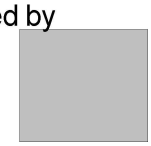
……

……

14.5 TESTING OF HYPOTHESIS IN CASE OF SMALL SAMPLES (t-Test)

We have already discussed that in case of large samples (i.e. $n \geq 30$,) the sampling distribution of the statistic is approximately normal. But in many decision making situations, due to time, money and other constraints, sometimes a researcher may have to take small sample of size less than 30 (i.e. $n < 30$,). In case of small samples also, the researcher can apply standard normal test for a specified population mean or equality of two population means provided the sample is drawn from a normal population where S.D. σ is known. But if σ remains unknown then standard normal test fails and in all such cases the sampling distribution turns out to be the student's 't' distribution. Sir William Gosset of Ireland in early 1900, under his pen name 'Student', developed a method for testing of hypothesis known as 't-test'. It is to be noted that the concept of degrees of freedom is associated with t-test. The graph of the 't' distribution approaches the normal curve as the degrees of freedom increase.

If $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ be a random sample of size drawn from a normal population with mean μ and variance σ^2, then the Student's 't' statistic is defined by



where $\bar{x}$ = sample mean

and 

is an unbiased estimator of σ^2. If s^2 be the sample variance given by



Then $ns^2 = (n-1)S^2$

Thus we have, 



Here, 't' follows t-distribution with (n-1) d.f.

14.5.1 Characteristics of t-Distribution

The characteristics of t distribution are given below:

(1) It is an exact distribution.

(2) The distribution is unimodal with

Mean=Median=Mode=0

(3) The t-distribution is less peaked than normal distribution at the centre and higher at both the ends.

(4) The shape of the -distribution depends on depends on the sample size . As increases, it approaches the normal distribution.

(5) The limits of t is $-\infty < t < \infty$.

(6) The critical values of at level of significance vary with the degrees of freedom.

14.5.2 Applications of t-Distribution

The -distribution has a number of applications in Statistics some of which are given below

(i) It is used to test the significance of single mean, population standard deviation being unknown.

(ii) It is used to test the significant difference between two population means under the assumption that population standard deviations being equal but unknown.

(iii) It is used to test the significance of an observed sample correlation coefficient.

Assumptions for Student's -test: The following assumptions are made in Student's -test:

(i) The parent population from which the sample is drawn is normal.

(ii) The drawn sample is random.

(iii) The population standard deviation σ remains unknown.

14.5.3 Hypothesis Testing for Single Population Mean

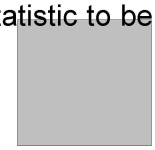
Suppose a random sample of size $n(n<30)$ with mean $\bar{x}$ is drawn from a population μ. We want to test the hypothesis that the population has the specified value μ_0 when the population S.D. σ is unknown.

Here we set up the null hypothesis $H_0 : \mu = \mu_0$.

Now depending upon the situation, against this hypothesis we may have any one of the following three alternative hypotheses

(i) $H_1 : \mu \neq \mu_0$ (Two-tailed test)
(ii) $H_1 : \mu > \mu_0$ (Right-tailed test)
(iii) $H_1 : \mu < \mu_0$ (Left-tailed test)

Since the sample is small, therefore under the null hypothesis the test statistic to be constructed is

 follows t distribution with $\nu = (n-1)$ d f.

where $s =$ sample S.D., $\mu = \mu_0$, σ is the population standard deviation, n is the sample size and $\bar{x}$ is the sample mean.

Note: For large sample ($n \geq 30$), we have

S.E. ($\bar{x}$) = 

On the other hand, for small sample ($n < 30$),

S.E. ($\bar{x}$) = 

The critical values (i.e., tabulated values) of t distribution given in the table for 't' in the appendix are the critical values of for right-tailed test. The corresponding values for left-tailed test will just be the negative of these values. If the values of for a two-tailed test at level are to be determined then instead of α, we are to take $\alpha/2$.

If the alternative hypothesis is two-tailed as defined in (1), then the

null hypothesis is to be rejected if the calculated value of $|t| > t_{\alpha/2}$.

Again if the alternative hypothesis is one-tailed as defined in (2) or (3), the null hypothesis is to be rejected if the calculated value of $|t| > t_{\alpha}$.

Example 10: Royal Tyres has launched a new brand of tyres for tractors and claims that under normal circumstances the average life of the Tyres is 40,000 km. A retailer wants to test this claim and has taken a random sample of 8 Tyres that yields mean of 39,750 with a standard deviation of 2618.61. He tests the life of the Tyres under normal circumstance. (Use $\alpha = 0.05$)

Solution: Here the size of the sample $(i.e., n)$ is 8 (< 30) and the population standard deviation is unknown, so test can be applied to test the hypothesis.

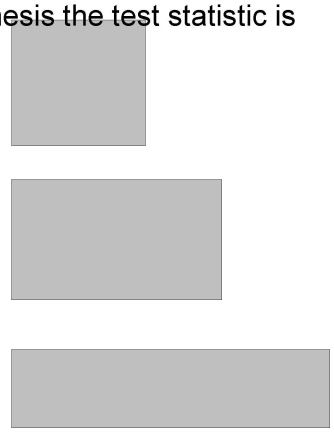
Given $\bar{x} = 39750$ and $s = 2618.61$

Now, we set up the null and alternative hypotheses as:

Null hypothesis $H_0 : \mu = 40{,}000$

Alternative hypothesis $H_1 : \mu \neq 40{,}000$

Under the null hypothesis the test statistic is



$$\therefore \quad |t| = |-0.233| = 0.233$$

Degrees of freedom= $n - 1 = 7$

The critical value of $|t_{0.05}(7)|$ for two-tailed test = 2.365

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 7 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the company's claim is valid under normal conditions.

Example 11: The mean price of shares of a company during the year 1992 was Rs. 85. In 1993, the mean price of shares of 20 randomly selected days is found to be Rs. 75 with a standard deviation of Rs. 10. Is there a significant difference in the prices of shares for the two years? (Given $|t_{0.05}(19)| = 2.093$)

Solution: Here the size of the sample $(i.e., n)$ is 20 (< 30) and the population standard deviation σ is unknown, so t test can be applied to test the hypothesis.

Given $\bar{x} = RS.75$ and $s = Rs.10$

Now, we set up the null and alternative hypotheses as:

Null hypothesis $H_0 : \mu = Rs.85$

Alternative hypothesis $H_0 : \mu \neq Rs.85$

Under the null hypothesis the test statistic is



Which follows t distribution with $(n-1)$ d f.

Now 



$\therefore \quad |t| = |-4.36| = 4.36$

Degrees of freedom = $20 - 1 = 19$

The critical value of $|t_{0.05}(19)|$ for two-tailed test = 2.093

Since the critical value of $|t|$ is greater than the critical value of $|t|$ at 5% level for 19 d.f., therefore we reject the null hypothesis at 5% level of significance and conclude that there is significant difference between the mean prices of shares in the two years.

Example 12: A fertilizer mixing machine is set to give 12 kg. of nitrate for every 100 kg. of fertilizer. Ten bags of 100 kg. each are examined. The percentage of nitrate so obtained is: 11, 14, 13, 12, 13, 12, 13, 14,11 and 12. Is there reason to believe that the machine is defective? (Use $\alpha = 0.05$)

Solution: Given $n = 10$

Let us consider the null and alternative hypotheses as

$$H_0 : \mu = 12 \quad \text{and} \quad H_1 : \mu \neq 12$$

We now calculate mean and standard deviation of the given sample as shown below:

Calculations of sample mean and standard deviation:

Variable x	$x - \bar{x} = x - 12.5$	$(x - \bar{x})^2$
11	-1.5	2.25
14	1.5	2.25
13	0.5	.25
12	-0.5	.25
13	0.5	.25
12	-0.5	.25
13	0.5	.25
14	1.5	2.25
11	-1.5	2.25
12	-0.5	.25
125	0	10.5

 and 

Under null hypothesis the test statistic is



Which follows t distribution with $(n-1)$ d.f

Now 





Degrees of freedom= $10-1=9$

The critical value of $|t_{0.05}(19)|$ for two-tailed test = 2.262

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 9 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the manufacturer's claim is valid, i.e., the machine is not defective.

14.5.4 Hypothesis Testing for Difference between Two Population Means

In case of two populations, our concern is to test if two independent samples have been drawn from two normal populations having the same means, the population standard deviations being equal. Let $\overline{x_1}$ and $\overline{x_2}$ be the means of two independent random samples of sizes n_1 and n_2 ($n_1 < 30, n_2 < 30$) drawn from two normal populations having means μ_1 and μ_2 respectively with equal but unknown standard deviations σ_1 and σ_2.

Thus the null hypothesis is $H_0 : \mu_1 = \mu_2$, i.e., samples have been drawn from the normal populations with same means.

The alternative hypothesis H_1 may be any one of the following depending upon the situation.

(1) $H_1 : \mu_1 \neq \mu_2$ (two-tailed)

(2) $H_1 : \mu_1 < \mu_2$ (left-tailed)

(3) $.H_1 : \mu_1 > \mu_2$ (right-tailed)

Assuming the null hypothesis to be true, the test statistic t is given by



i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f

where 

and 



is an unbiased estimate of the common population variance σ^2 based on both the samples.

The decision to reject or accept the null hypothesis is similar to that used in one-sample tests.

Example 13: The IQs (intelligence quotients) of 16 students from one area of a city showed a mean of 107 with a standard deviation of 10, while the IQs of 14 students from another area of the city showed a mean of 112 with a standard deviation of 8. Is there a significant difference between the IQs of the two groups at 5% level of significance?

Solution:

Given $n_1 = 16, \quad n_2 = 14, \quad \overline{x_1} = 107, \quad \overline{x_2} = 112, \quad s_1 = 10$ and $s_2 = 8$

We set up null and alternative hypothesis as

Null hypothesis is $H_0 : \mu_1 = \mu_2$

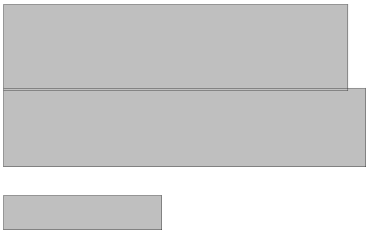
Alternative hypothesis is $H_0 : \mu_1 \neq \mu_2$

Since sizes of both the samples are small and population standard deviations are unknown therefore to test the hypothesis we apply t test given by



i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f.

Now



Thus $S = 9.44$



The degrees of freedom = $n_1+n_2-2=28$

The critical value of $|t_{0.05}(28)|$ for two-tailed test = 2.02

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 28 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the difference between the two population means is not significant.

Example 14: In a test given to two groups of students, the marks obtained are as given below:

First group : 25 32 30 32 24 14 32
Second group: 24 34 22 30 42 31 40 30 32 35

Examine the significance of the difference between the arithmetic mean of the marks secured by the students of the above two groups.

Solution: Here we set up null and alternative hypotheses as

Null hypothesis is $H_0 : \mu_1 = \mu_2$ i.e., there is no significant difference between mean marks secured by students of the two groups.

Alternative hypothesis is $H_0 : \mu_1 \neq \mu_2$

Since sizes of both the samples are small and population standard deviations are unknown therefore to test the hypothesis we apply t test given by



i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f.

Now



Calculations of $\overline{x_1}$, $\overline{x_2}$ and pooled sample standard deviations are given below:

Calculation of $\overline{x_1}$, $\overline{x_2}$ and S

x_1	$x_1 - \overline{x_1}$ $= x_1 - 27$	$(x_1 - \overline{x_1})^2$	x_2	$x_2 - \overline{x_2}$ $= x_2 - 32$	$(x_2 - \overline{x_2})^2$
25	-2	4	24	-8	64
32	5	25	34	2	4
30	3	9	22	-10	100
32	5	25	30	2	4
24	-3	9	42	10	100
14	-13	169	31	-1	1
32	5	25	40	8	64
			30	-2	4
			32	0	0
			35	3	9
189		266	320		350

Here 

Sample variance, 

Sample variance, [redacted]

Now, [redacted]

[redacted]

[redacted]

[redacted]

The degrees of freedom = $n_1+n_2-2=15$

The critical value of $|t_{0.05}(15)|$ for two-tailed test = 2.131

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 28 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the mean marks obtained by the students of two groups do not differ significantly.

Example 15: Two salesmen, A and B, are employed by a company. To study if there is any significant difference between the average sales of the two salesmen, the HRD manager of the company conducted a survey yielding the following information:

	Salesman A	Salesman B
Number of sales	20	22
Average weekly sales (Rs. lakh)	30	25
Standard deviation (Rs lakh)	10	7

Solution: First we set up the null hypothesis that there is no significant difference between the average sales of the two salesmen. That is, $H_0 : \mu_1 = \mu_2$

The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$

Since sizes of both the samples are small and population standard deviations are unknown therefore to test the hypothesis we apply t test given by

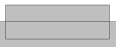


i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f.

Now





Thus 

The degrees of freedom = $n_1+n_2-2=40$

The critical value of $|t_{0.05}(40)|$ for two-tailed test = 2.021

Since the calculated value of $|t|$ is less than the critical value of $|t|$ at 5% level for 40 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the average sales by the two salesmen are not significantly different.



CHECK YOUR PROGRESS

Q 6: To the following information (a) State null and alternative hypotheses, (b) State the decision rule, (c) Compute the pooled estimate of the population variance, (d) Compute the test statistic. (e) What is your decision regarding H_0 ?

A random sample of 10 observations from one population revealed a sample mean of 23 and a S.D. of 4. A random sample of 8 observations from another population revealed a sample mean of

26 and a S.D. of 5. Use 5% level of significance to test if there is any significant difference between population means?

14.5.5 Paired $t-$Test for Difference between Two Population Means

In the $t-$test used for difference of means, the two samples were independent of each other whereas in case of paired $t-$test we have

(i) The sample sizes (n_1 and n_2) are equal.

(ii) The sample observations $(x_1, x_{2,}.....,x_n)$ and $(y_1, y_2,....., y_n)$ are not completely independent but they are dependent in pairs i.e., the pairs $(x_1, y_1),(x_2, y_2),.....(x_n, y_n)$ correspond to the 1st, 2nd, ……..,nth unit respectively.

For example, the management of a company plagued by poor productivity realizes the need to provide technical training to employees. It hires a researcher to measure the productivity levels of a sample of 25 employees. The productivity levels are measured again after a one-month technical training programme. In this kind of pre-and post-training study, which are selected before and after the study cannot be treated as independent because each observation in sample 1 is related to the observation in sample 2. The productivity scores obtained before training is related to the scores obtained after training because the two measurements are apply to the same person.

Now we explain the paired $t-$ test for difference between the means of two related populations as follows:

Suppose a business concern is interested to know if advertisement is really effective in promoting sales of a particular product. Let $(x_1, x_{2,}.....,x_n)$ be a sample of sales of the product in departmental stores for a certain period before advertisement campaign and $(y_1, y_2,....., y_n)$ be the sample of corresponding

sales of the same product and for period in the same departmental stores respectively. Now $(x_i, y_i), i = 1,2,.....,n$ is the pair of sales in the i[th] departmental store before and after advertisement. Suppose we want to test the significance of the difference between the sales, and.

Let $d_i = x_i - y_i, (i = 1,2,.....,n)$

We set up the null hypothesis that there is no significant change in sales after advertisement campaign. i.e., $H_0 : \mu_1 = \mu_2$.

The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$.

Under the null hypothesis, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{n-1}$$

where

and

Example 16: An electronic company arranged a special training programme for one segment of its employees. In order to measure the change in the attitude of its employees after the training, the company has prepared a well-designed questionnaire consisting of 10 questions on a 1 to 5 rating scale (1 is strongly disagree and 5 is strongly agree). The company selected a random sample of 10 employees. The scores are given below:

Employees	Scores before training	Scores after training
1	25	32
2	26	30
3	28	32
4	22	34
5	20	32
6	30	28
7	22	25
8	20	30
9	21	25
10	24	28

Test whether there is a significant change in the attitude of employees after the training programme. (Use $\alpha = 0.10$)

Solution: Here, scores before training (x) and after training (y) are not independent but are paired together. Hence, to test the difference between the means of two related populations we apply the paired $t-$test.

We set up the null and alternative hypotheses as:

The null hypothesis $H_0 : \mu_1 = \mu_2$ i.e., there is no significant effect of the training.

The alternative hypothesis $H_1 : \mu_1 \neq \mu_2$.

Under the null hypothesis, the test statistic is



where 

and 

Calculations of $\bar{d}$ and S^2 are given below:

Employees	Scores before training (x)	Scores after training (y)	$d = x - y$	d^2
1	25	32	-7	49
2	26	30	-4	16
3	28	32	-4	16
4	22	34	-12	144
5	20	32	-12	144
6	30	28	2	4
7	22	25	-3	9
8	20	30	-10	100
9	21	25	-4	16
10	24	28	-4	16
Total			-58	514



Thus $S = \sqrt{19.733333} = 4.44$

$|t| = 4.13$

The degrees of freedom = 9

The critical value of $|t_{0.10}(9)|$ for two-tailed test = 1.833

Since the calculated value of $|t|$ is greater than the critical value of $|t|$ at 10% level for 9 d.f., therefore we reject the null hypothesis at 10% level of significance and conclude that there is a significant difference in the attitude of employees before and after the training.



CHECK YOUR PROGRESS

Q 7: To the following (a) State null hypothesis, (b) State the decision rule, (c) Compute the test statistic. (e) What is your decision regarding H_0?

An IQ test was administered to 5 persons before and after they were trained. The results are given below:

Candidate:	I	II	III	IV	V
IQ before training:	110	120	123	132	125
IQ after training:	120	118	125	136	121

Test if there is any change in IQ level after the training programme.

.....

.....

.....

.....

.....

.....



14.6 LET US SUM UP

- Statistical tests that do not require any prior information about the population are termed as non-parametric tests. This unit focuses on only χ^2 (Chi-square) distribution and the related χ^2 test. χ^2 distribution is the family of curves with each distribution being defined by the degree of freedom associated to it.

- χ^2 test can be used for a variety of purposes. χ^2 test provides a platform that can be used to ascertain whether the theoretical probability distribution coincides with the empirical sample distribution. This is commonly known as χ^2 goodness of fit test. It can also be used to test the independence of independence of two variables. χ^2 test of independence makes use of the contingency table to determine the independence of two variables. This test can also be used for estimating the population variance. Again the test of homogeneity is applied to determine whether two or more populations are homogeneous with respect to some characteristics of interest.

- The test can be applied for dependent samples and the testing procedure is termed as "paired t test or t test for related samples." Under this approach, observations in 1st sample are related to the observations in 2nd sample.

14.7 FURTHER READING

1) Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.
2) Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.
3) Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.

14.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) T, (ii) F, (iii) T, (iv) T, (v) T.

Ans to Q No 2: (a) , (b) The critical value of at 5% level of significance for 3 degrees of freedom i.e., $\chi^2_{0.05,3} = 7.8147$

Ans to Q No 3: $d.f. = 3 - 1 = 2$

The critical value of at 5% level of significance for 2 degrees of freedom i.e., $\chi^2_{0.05,2} = 5.9915$

(a) The decision rule is to reject if

(b)

(c)Reject H_0.

Ans to Q No 4: (a) Contingency table.

(b) Null hypothesis is H_0 : There is no relationship between income and whether the person played the lottery.

Alternative hypothesis is H_1 : There is a relationship between income and whether the person played the lottery.

(c) The degrees of freedom, $d.f. = (2-1)\times(3-1) = 2$

(d) Reject H_0 if $\chi^2 > 5.9915$,The critical value of χ^2 at 5% level of significance for 2 degrees of freedom i.e., $\chi^2_{0.05,2} = 5.9915$

(e)

= 6.544

(f) Reject H_0.

Ans to Q No 5: Chi-square test of homogeneity is useful in a situation when we intend to verify whether several populations are homogeneous with respect to some characteristic of interest.

Ans to Q No 6: (a) Null hypothesis is $H_0 : \mu_1 = \mu_2$, Alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$,

(b) Here $d.f. = 10 + 8 - 2 = 16$, Therefore tabulated value of $|t_{0.05}(16)|$ for two-tailed test = 2.120

The decision rule is to reject H_0 if $|t| > 2.120$.

(c) Pooled estimate of the population variance,



(d)



(e) H_0 may be accepted.

Ans to Q No 7: (a) Null hypothesis is H_0 : There is no change in IQ level after training programme.

(b) Reject H_0 if $t > 4.6$, the tabulated value of $|t_{0.05}(4)|$ for two-tailed test.

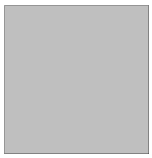
(c)



Therefore,



(d) H_0 may be accepted.



14.9 MODEL QUESTIONS

Q 1: Define Chi-square distribution. Discuss the important properties of the distribution.

Q 2: Under what situations Chi-square test can be applied?

Q 3: Write down the applications of Chi-square test.

Q 4: The figures given below are (a) the theoretical frequencies of a distribution, and (b) the frequencies of the normal distribution having

the same mean, standard deviation and the total frequency as in (a).

(a)	1	5	20	28	42	22	15	5	2
(b)	1	6	18	25	40	25	18	6	1

Apply χ^2 -test of goodness of fit.

[Hints: Since the observed and the expected frequencies are less than 5 at the beginning and at end of the series, these frequencies are grouped

Degrees of freedom = 4, Tabulated value of $\chi^2_{0.05}(4) = 9.49$

H_0 may be accepted implying that the fit is good.]

Q 5: A Mumbai film director claims that his films are liked equally by males and females. An opinion survey of a random sample of 1000 film-goers revealed the following results:

	Liked	Disliked
Male	402	193
Female	245	160

Is the film director's claim supported by data?

[Given : $\chi^2_{0.05}$ = 3.84 5.99 7.81

d.f.= 1 2 3]

Q 6: A certain oil corporation conducts a seismic test on land to assess the possibility of extraction of oil from the land. From the past 200 records, the following test results and oil yields were obtained:

Test results	Oil yield frequencies		
	High	Medium	Low
Good	44	22	13
Fair	15	19	26
Bad	5	8	47

Can we say, at 1% level of significance that oil yields and test results are associated?

Q 7: The following table gives the classification of 100 workers according to sex and the nature of work. Test whether the nature of work is independent of the sex of the work.

	Skilled	Unskilled
Male	40	20
Female	10	30

Q 8: A certain drug was administered to 456 males out of a total of 720 in a certain locality to test its efficiency against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease. (The table value of χ^2 for 1 degree of freedom at 5% level of significance is 3.84)

	Infection	No infection	Total
Administering the drug	144	312	456
Without administering the drug	192	72	264
Total	336	384	720

Q 9: A sample of 20 observations gave a standard deviation of 3.72. Is this compatible with the hypothesis that the sample is drawn from a normal population with variance 4.35?

[Hints: H_0 : The sample is drawn from a normal population.
Under H_0, the test statistic $\chi^2 = \frac{ns^2}{\sigma^2}$ follows χ^2 distribution with $n-1 = 20-1 = 19\ d.f.$ Table value of $\chi^2_{0.05}(19) = 30.144.$ is rejected at 5% level.]

Q 10: The sales data of an item in six shops before and after a special promotional campaign are as under:

Shops :	A	B	C	D	E	F
Before campaign :	35	28	31	48	50	42
After campaign :	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5% level of significance?

(Significant values of $|t|$ at 1% level and 5% level for 5 d.f. respectively are 2.015 and 2.571)



REFERENCES (For All Units of the Course)

1) Feller, W. (1968). *An Introduction to Probability Theory and its Application*, Vol I, Wiley Eastern.

2) Gujrati, D. (1995). *Basic Econometrics*. McGraw-Hill, Inc., Singapore

3) Gun, A. M., Gupta, M K., & Dasgupta, B. (2002). *Fundamentals of Statistics*, Volume One; World Press Private Ltd. Kolkata

4) Gupta, S. C. (1994). *Fundamentals of Statistics*. Himalayan Publishing House, New Delhi.

5) Gupta, S. P. (1995). *Statistical Methods*. Sultan Chand & Sons, New Delhi.

6) Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.

7) Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.

8) Medhi, J. (2000). *Statistical Methods: An Introductory Text*, New Age International (P) Limited, Publishers.

9) Mukhopadhyay, P. (1996), *Mathematical Statistics*. New Central Book Agency (P) Ltd.

10) Mukhopadhyay, P. (1991). *Theory of Probability*. New Central Book Agency (P) Ltd.

11) Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.

Name of the Paper: .....
Course Code: .....

Centre for Internal Quality Assurance (CIQA)
Krishna Kanta Handiqui State Open University
City Office: Housefed Complex, Guwahati-781006

Learner's Feedback on Course

Dear Learner,

Regarding the course as mentioned above, we would like to know your opinions and comments so as to improve the quality of self learning materials in future. Please respond to the following statements by ticking the number you feel most reflect your opinion. After completion of the additional comments, please detach the page and send/mail the same to us at the address given below.

The Director, Centre for Internal Quality Assurance, KKHandiqui State Open University
Housefed Complex, Dispur, Guwahati-781006
(E-mail id: ciqa@kkhsou.in)

1) Approximately how many hours did you spend for studying the units in the course? ☐☐☐

2) Please give your opinions (by ✓ mark) to the following items based on your reading of the block:

Sl. No.	Statements	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I)	The SLMs of the course fulfil the learning objectives set out in the beginning of each unit					
II)	The units of the SLM could satisfy my academic needs and expectations					
III)	The Unit writers have excellent knowledge about the course contents					
IV)	Language and contents of the units were presented at a level which I could readily understand					
V)	Ample opportunity for participation in the activities provided in the units					
VI)	Used enough Illustrations (Diagrams, tables etc.) for conceptual clarity					
VII)	Quality of content is engaging, relevant, and up-to-date					
VIII)	The self check questions are very helpful					
IX)	The Possible/model questions and the answers to check my progress have benefited me a lot					

Additional Comments: *(Please feel free to provide your open comments)*

1) Which aspects of the SLM, according to you, worked well?

.....

.....

2) What sort of changes/improvements do you feel KKHSOU could implement to improve the overall quality of the SLM?

.....

.....

Thank you for taking the time to complete this form.